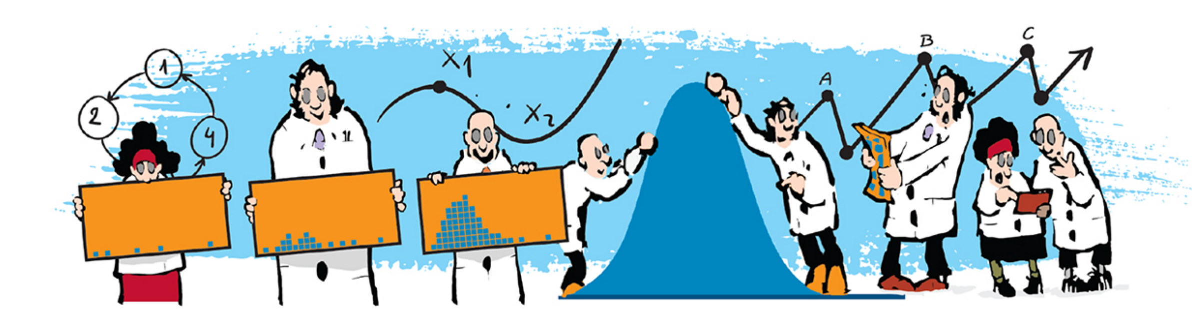
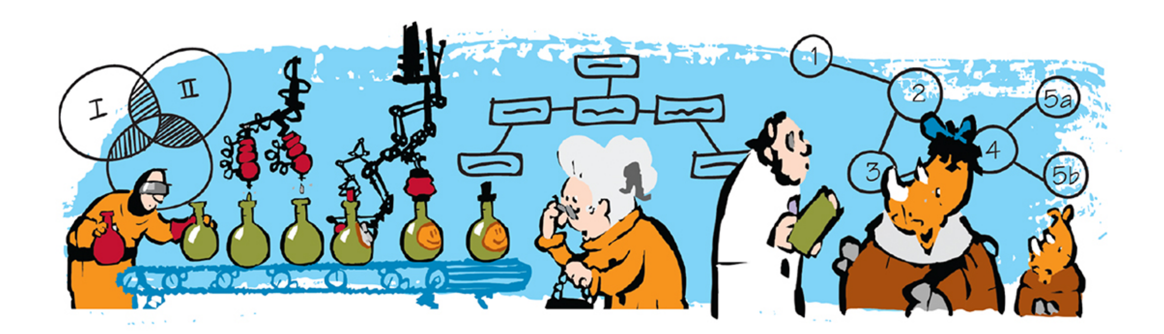
Грейди Клейн и Алан Дебни

СТАТИСТИКА



БАЗОВЫЙ КУРС



В КОМИКСАХ



СТАТИСТИКА

Базовый курс в комиксах

Грейди Клейн и Алан Дебни



Перевод с английского Ольги Терентьевой

Москва
«Манн, Иванов и Фербер»
2017

УДК 311.1
ББК 65.051
К48

Научный редактор Ирина Николаева

Издано с разрешения
Synopsis Literary Agency c/o THE SYNOPSIS NOA LLP

На русском языке публикуется впервые

Клейн, Грейди
К48 Статистика. Базовый курс в комиксах / Грейди Клейн, Алан Дебни ; пер. с англ. О. Терентьевой ; [науч. ред. И. Николаева]. — М. : Манн, Иванов и Фербер, 2017. — 240 с.

ISBN 978-5-00100-260-4

Не только полезный, но и веселый курс базовой статистики. Автор и иллюстратор объясняют сложные понятия на простых и забавных примерах, доказывая, что статистика — вокруг нас.

Прочитав эту книгу, вы научитесь собирать данные, делать выборки и проверять гипотезы по любой проблеме — будь то решение о покупке новой машины или подсчет степени взаимной ненависти жителей враждующих планет. Теперь вас не введут в заблуждение показатели средних зарплат по галактике и предвыборные рейтинги, составленные на основе смещенного распределения. Вы узнаете, почему идеальная форма в статистике не менее важна, чем содержание. И в конце концов, получите ответ на важный вопрос, кого огры кидают дальше — эльфов или гномов.

Если же вы захотите мыслить и говорить как статистик, в конце книги вас ждет «Математическая пещера», богатая на формулы и детали.

Книга будет полезна всем, кто хочет познакомиться со статистикой и научиться анализировать данные.

УДК 311.1
ББК 65.051

ISBN 978-5-00100-260-4

Посвящается Анне, Лиаму и Бенджамину.
Г. К.

Посвящается Эллиотту, Луизе и Нику.
А. Д.

СОДЕРЖАНИЕ

Вступление

Она повсюду



Статистика
окружает
нас!

Большинство из нас так или иначе имеют дело со статистикой каждый день...





...даже если мы не жонглируем цифрами, зарабатывая себе на жизнь.

Статистику «излучают» наши телевизоры...



...она просачивается из телефонов...



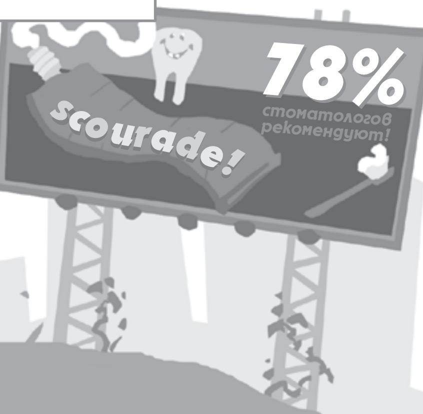
...льется из радиоприемников...



...и оставляет информационный мусор на нашем пути.







От нее не скрыться.

Статистика повсюду:









Статистика с нами с самого рождения...



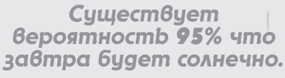
...и нравится нам это или нет, но мы и сами пополним статистику, когда умрем.



К счастью, всему этому есть хорошее объяснение.











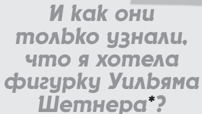




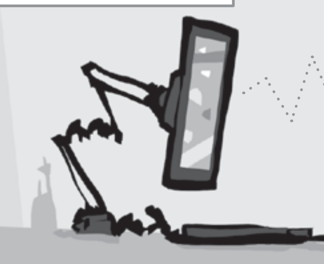
























* Уильям Шетнер — канадский актер, известный по роли капитана звездолета Джеймса Тиберия Кирка в сериалах и фильмах «Звездный путь». Прим. ред.

Статистика помогает побеждать на выборах...

Всего 23% моих избирателей считают меня абсолютным болваном!

...и возводить электростанции...

Спровоцирует ли наша ядерная установка мутации у местных жителей?

...и зарабатывать деньги...

При сохранении нынешнего состояния рынка...

...уже завтра я буду на 12–15% богаче!

...и доказывать свое превосходство...

Ха, на моем счету больше хоум-ранов, чем у тебя.

Да ты использовал стероиды, и у меня есть статистические данные, подтверждающие это.

Так что же делает статистику такой невероятно полезной?



Самое простое **объяснение** заключается в том, что статистика помогает контролировать **огромное количество важных вещей...**







...что, в свою очередь, помогает **лучше понять наш сложно устроенный мир...**

...и **управлять им.**



Но **настоящая сила** статистики **все же в другом.**

Вот в чем кроется **истинная причина** того, почему всем нужна статистика.



Представьте себе, что мы хотим узнать **средний вес...**

...всей рыбы в озере.



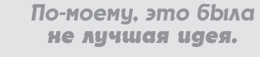
Если бы мы **осушили озеро и взвесили каждую рыбку...**

...то **получили бы всю необходимую информацию** и высчитали средний вес.





Но по очевидным причинам мы **не можем этого сделать.**





С другой стороны, если мы поймаем **100 рыбешек и взвесим их...**

Эти **100** рыбок весят **112** кг.

Следовательно, **в среднем одна рыбешка** весит **1,12** кг!

...мы получим **неполную информацию** о всей рыбе в озере.

Итак, теперь нам известен средний вес рыбы **в этой выборке.**

...но мы по-прежнему не знаем средний вес **остальной рыбы в озере.**

Но вот что **интересно:**

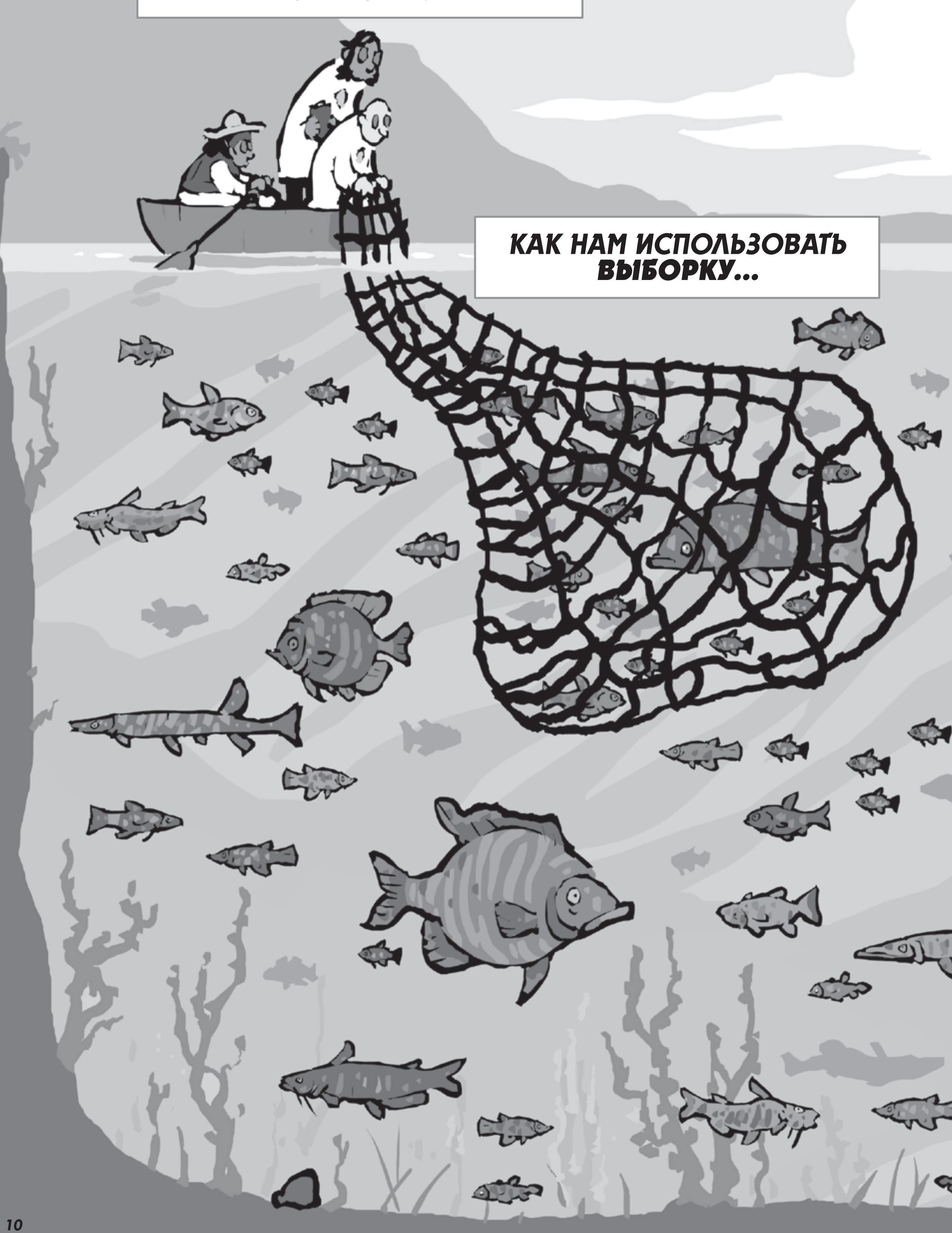
прибегнув к инструментарию статистики, мы можем **использовать эту неполную информацию...**

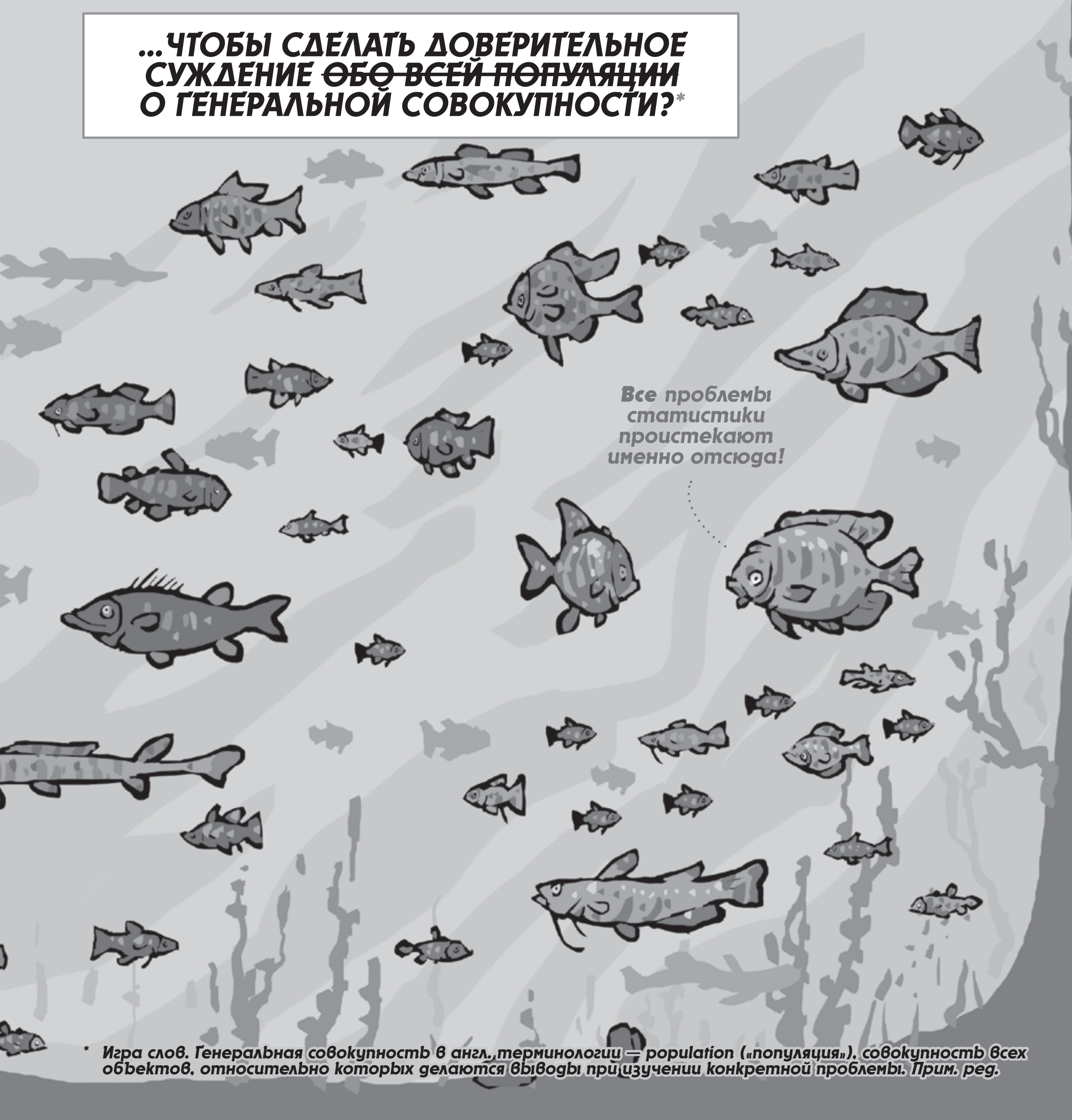
Статистика предполагает **использование той рыбы, которую мы поймали...**

...чтобы сделать **доверительное суждение** относительно **всей рыбы** в этом озере.






Эта книга отвечает
на **фундаментальный** вопрос
статистики:
КАК НАМ ИСПОЛЬЗОВАТЬ
ВЫБОРКУ...



* Игра слов. Генеральная совокупность в англ. терминологии — population («популяция»), совокупность всех объектов, относительно которых делаются выводы при изучении конкретной проблемы. Прим. ред.



...и **изучать** ее.



А затем, во второй части, мы научимся использовать выборку, чтобы
получить качественные результаты
для генеральной совокупности...

...используя процесс, который носит название
«статистическое заключение».



Таким образом мы сможем обработать **большие объемы данных...**



...высчитать **доверительные интервалы...**



...и **проверить гипотезы.**



И в более общем смысле мы получим представление о том, что **можно...**

...и что **нельзя...**

...делать с помощью статистики.









В этой книге мы **сфокусируемся** на **основных понятиях.**



Таких, как стандартные отклонения...

...и распределение выборки...

...и вероятности...

...И ДОСТОВЕРНОСТЬ!





Если же вам интересно узнать о **деталях...**

Например, что, черт возьми, означают все эти формулы и символы?



...то вы найдете их в приложении, которое называется **«Математическая пещера».**



Часть первая

СБОР СТАТИСТИЧЕСКИХ ДАННЫХ



Глава 1

ЧИСЛА

В этом углу ринга боксер, который весит



50,8 триллиона нанограммов...

...КАРЛИК!

А в этом углу боксер, который весит

0,193 тонны...



...ГИГАНТ!

Как мы узнали из предисловия, статистика — это не только цифры.

Доброе утро, это оператор 3810448, чем я могу вам помочь?

Статистика нужна, чтобы измерить нашу уверенность в чем-либо.

Как бы то ни было, статистика в действительности — это упорная борьба с цифрами...

С кем бы вы хотели сразиться?

На моем счету 147 побед и всего 17 поражений.

А я побеждаю в 89,6% случаев!




Некоторые числа большие...
В вашем мозгу 10^{14} нервных соединений!
То есть 100 000 000 000 000 соединений!


...а некоторые маленькие.
Каждый атом в 10^{-17} раз меньше глазного яблока.
То есть в 100 000 000 000 000 000 раз!


Но некоторые большие числа описывают совсем маленькие вещи...
В Техасе обитает более 500 млрд насекомых.


...и наоборот.
В Солнечной системе всего одна звезда.


Некоторые числа говорят о чем-то хорошем...


Мы выиграли, забив на два гола больше, чем соперники!


..а некоторые — о плохом.
Индекс Доу-Джонса упал на 423 пункта! Ужас!


Но бывает и так, что какие-то положительные числа описывают отрицательные вещи...
Безработица в прошлом месяце увеличилась на 4%.


...и наоборот.
Количество убийств в городе снизилось.
Но и это еще не все...







...другие кажутся **обнадеживающими.**





Некоторые говорят о **серьезных достижениях...**



...другие **не о столь серьезных...**







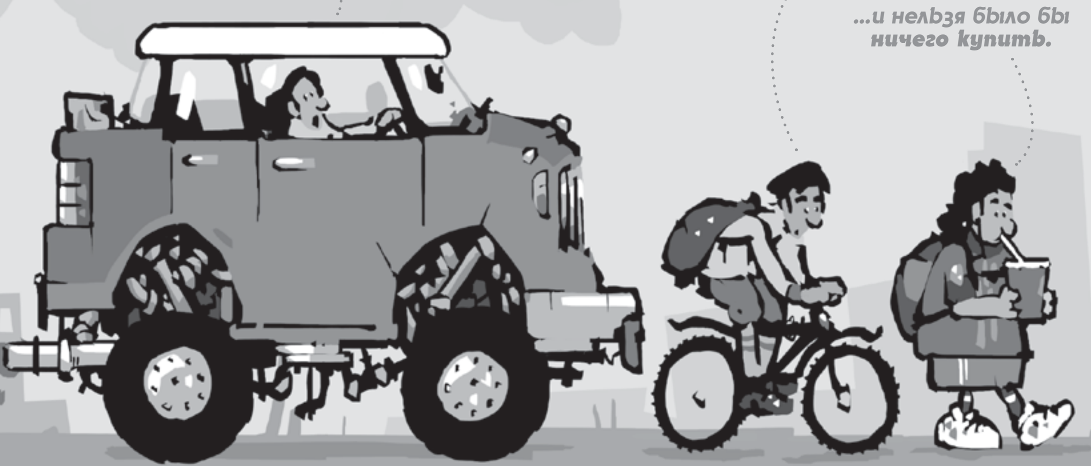
Все эти **факты позволяют** с легкостью **использовать числа...**

...чтобы кого-нибудь обмануть.



К сожалению, это может вынудить людей относиться к цифрам с недоверием...

...и не ценить их истинную силу.




Решение проблемы в том...
Некоторые цифры действительно отображают положение дел.
Но нужно помнить, что бывает и неверная информация!
Как же тогда понять, что правда, а что ложь?


...чтобы относиться ко всем числам...
...независимо от того, большие они...
...или маленькие...
...или и вовсе вгоняют в сон...
Хррр...


...с долей здорового скептицизма.
Это печенье органическое на 100% и на 98,3% подходит веганам...
...а лактоово-показатель находится в пределах рекомендованной нормы.
Пожалуй, возьму печеньку, но с долей здорового скептицизма.
Это первый урок нашей книги.


Помогите!
Что-то я сомневаюсь в этих цифрах.
И хорошо!
Пусть это ощущение будет в радость!
В статистике это верный подход!


Дальше мы узнаем, как инструменты статистического наблюдения помогают нам делать с помощью цифр прогнозы на будущее.
Во Вселенной по меньшей мере 100 млрд черных дыр!
Я знаю, что вы не все из них видели своими глазами...
...как же вы можете быть уверены в этом?
Скажу пока так: у людей всегда есть причины жонглировать цифрами...
Эти цифры доказывают, что я прав!
А вот эти — что ты ошибаешься!
...и будет нелишним подумать, что это за причины.
В 98% случаев это лекарство эффективно для людей с вашим заболеванием...
...которое в 14,8% случаев заканчивается летальным исходом.
Зачем они говорят мне все это?
$





...или нет...



...сталкиваясь с ними, вы должны
задать себе несколько вопросов.



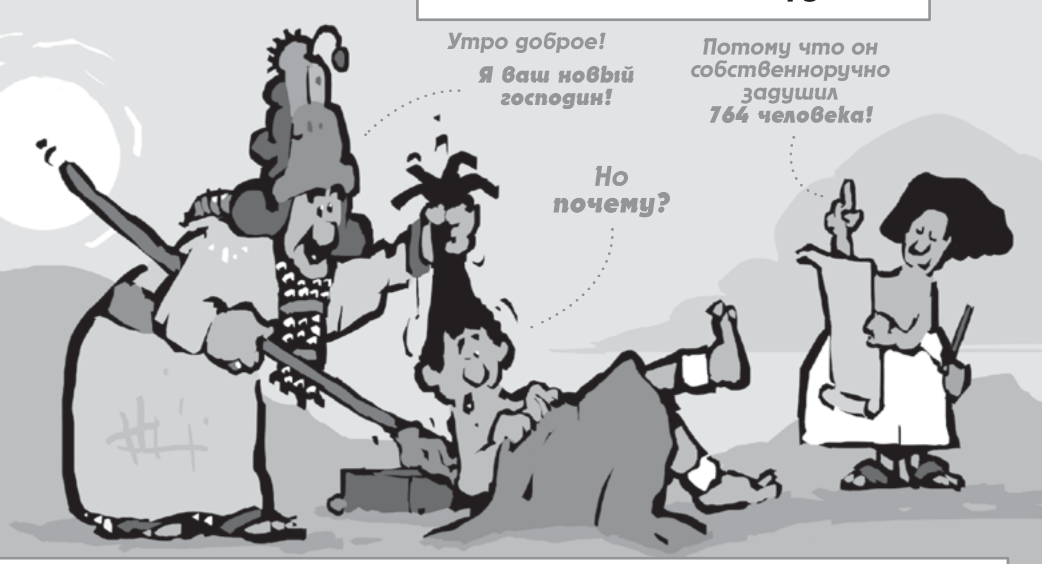
Глава 2

СЛУЧАЙНЫЕ СЫРЫЕ ДАННЫЕ



С момента **сотворения мира...**

...у людей есть **потребность считать все, что их окружает.**



И правда, самые ранние **формы письменности** были придуманы, чтобы вести **математические подсчеты.**



По мере развития **цивилизации...**

...появлялось **множество вещей,** которые нужно было считать.





*Иногда **невозможно** подсчитать все, что мы хотим.*



Вот поэтому когда-то давным-давно кому-то в голову пришла мысль о том, чтобы...

*...исследовать **выборку...***

*...и, изучив ее, сделать выводы о **генеральной совокупности.***

Я знаю, что делать!

Я поговорю с **отдельными** воинами...

...и использую это знание, чтобы сделать предположения...

...обо **всех** остальных!

***Использование выборки для описания генеральной совокупности — это* умно...**

***...но есть несколько нюансов, о которых следует помнить, прежде чем* браться за дело.**





Во-первых, руководствуясь данными выборки,* невозможно *судить с* абсолютной точностью *обо всей совокупности.







Именно поэтому статистика нужна для того, чтобы делать* максимально точные *предположения...

...а не быть абсолютно уверенным.







Во-вторых, если мы застопорились на единственной выборке...

Я могу сделать выводы обо всех кальмарах, живущих в океане...

...изучив только эти 35 особей!

...лучше убедиться, что мы собрали ее аккуратно!



Потому что любая ошибка, допущенная нами при определении выборки...

...может кардинально исказить наши выводы о генеральной совокупности.





Опрашивая

Подсчитывая

Взвешивая и замеряя

Пробуя на зуб

...и это далеко не простая работа.

Добиваться точности измерений бывает особенно сложно, если речь идет о крупном...

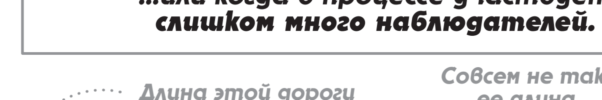
Возьмите-ка линейку...

...чтобы измерить тех драконов...

Это помогает сфокусироваться на деталях.

...или мелком масштабе...

...и вот этих стрекоз.

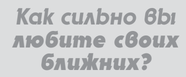





Определение выборки также затруднительно в тех случаях, когда мы пытаемся понять, о чем думают люди...


Какой цвет более выигрышный: красный или зеленый?


Я дальтоник.


...или что они чувствуют...


Как сильно вы любите своих ближних?


Да я их ненавижу!



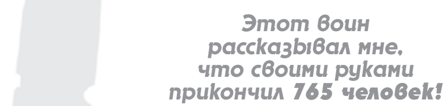


...или выспросить у них то, о чем они, возможно, даже не хотят говорить...


Сколько банков вы ограбили?




Зависит от того, кто интересуется.


...или когда люди преувеличивают.


Этот воин рассказывал мне, что своими руками прикончил 765 человек!


И ты ему поверил?





Возможно, **основная сложность** в формировании выборки...

Мне предстоит опросить **100 воинов!**

Но **кого именно** я должен **выбрать?**

...это понимание того, что **именно следует в нее включить.**

Кому отдать предпочтение? Может, этим **вежливым** парням, которые постоянно сидят в кофейне...

...или тем **устрашающего вида** громилам в спортзале?

Следите за тем, чтобы ваше мнение не было **предвзятым...**

...это может привести к **искажению данных о генеральной совокупности.**

Если вы опросите слишком много **вежливых** воинов...

...вам может показаться, что армия — **гораздо более приятное место, чем на самом деле!**

Если же вы опросите слишком много вояк **грозного вида...**

...вы решите, что в армии **страшнее, чем на самом деле!**

В идеале хотелось бы собрать такие данные, которые бы точно отражали генеральную совокупность.



*Это задание кажется мне просто **невыполнимым...***





*...на такой случай у статистиков припасен **надежный способ.***

Чтобы избежать предвзятого суждения, мы всегда делаем случайную выборку.

Не переживайте!

Просто закройте глаза!

И положитесь на волю случая!

Делать случайную выборку — это простая идея...

Давайте поместим всю армию целиком в этот шлем...

...и будем наугад вытаскивать по одному воину.

Это же как игра в «Бинго»!

...которая может на деле оказаться сложной.

ЗАБИРАЙТЕСЬ, ПАРНИ!

На практике нам нужно перебрать в уме все факторы, которые могут оказать влияние на нашу выборку...

Я не хочу включать в опрос воинов, спящих беспробудным сном в канаве...

...или тех, от которых плохо пахнет...

...или живущих в этом городе...

...и проследить за тем, чтобы это нам не помешало.





*Конечно, когда мы делаем **случайную** выборку...*







*...то по-прежнему не **гарантируем**, что она **даст нам представление** о генеральной совокупности **с зеркальной точностью**.*

*На самом деле любая **случайная выборка...***

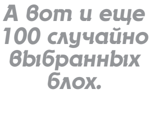
*...будет **наверняка отличаться от генеральной совокупности** в целом...*

*...равно как и от **любой другой** случайной выборки.*











*Почему же случайная выборка так хорошо работает? Потому что мы **можем** взять для рассмотрения как эту выборку...*

...так и любую другую...



...и если уж они разные...



*...это **ТОЛЬКО** по воле случая.*







Делать случайную выборку довольно сложно...

Я провожу исследование **всей** рыбы, живущей в море...

...и мне нужно **наугад** выбрать экземпляры для исследования.

Остановлюсь, пожалуй, вот на этой рыбешке, оказавшейся здесь **в этот момент.**

...но **очень важно** делать это правильно...



Иногда стоит дойти до края Земли...

...или нырнуть на морское дно!





* См. стр. 214.

В этой главе мы узнали, как случайная выборка способна помочь нам избежать необъективности.

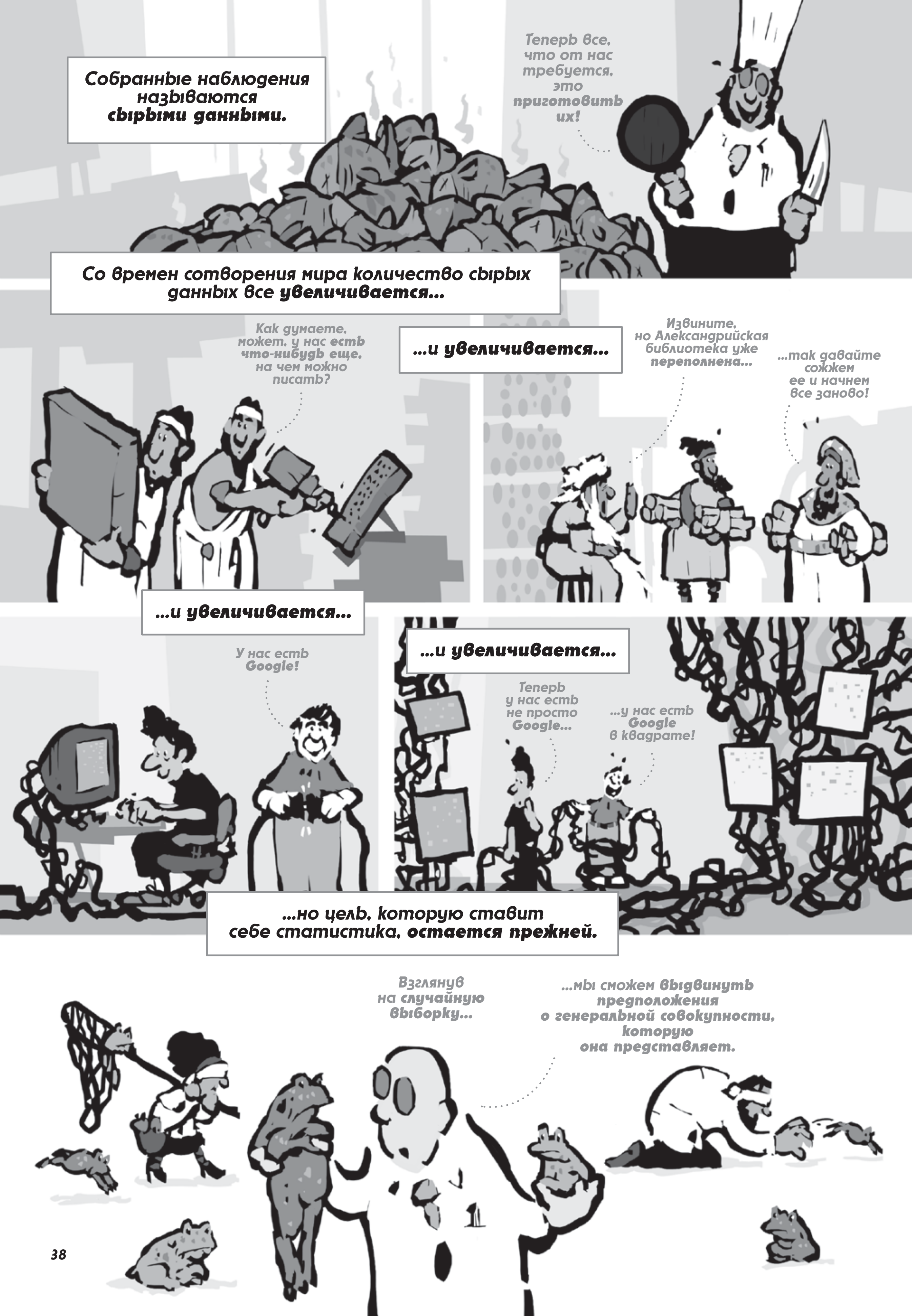




Но случайные выборки представляют собой существенную часть статистической системы, о которой мы узнаем позже.

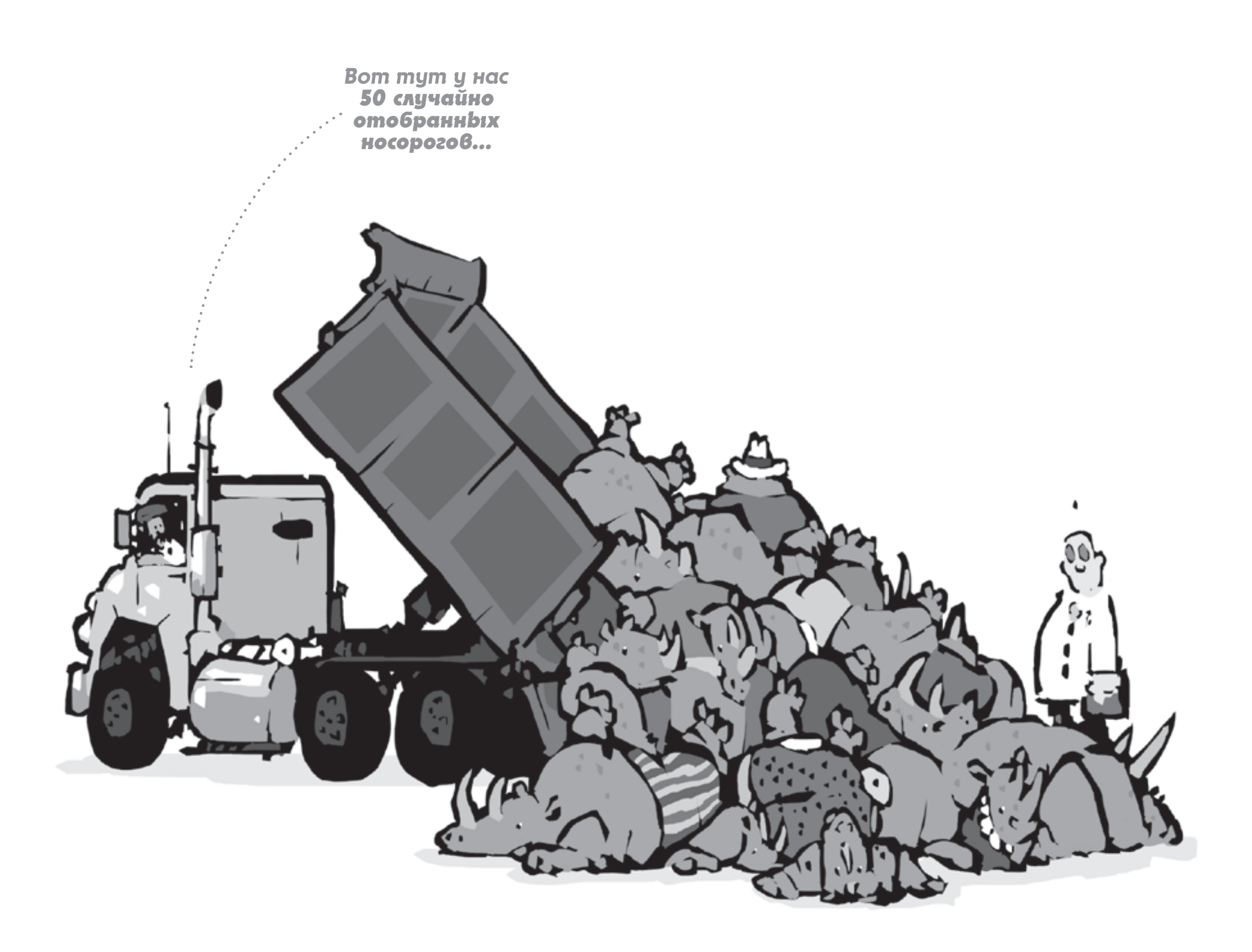





Собранные наблюдения называются **сырыми данными.**
Теперь все, что от нас требуется, это **приготовить** их!
Со времен сотворения мира количество сырых данных все **увеличивается...**
Как думаете, может, у нас есть что-нибудь еще, на чем можно писать?
...и **увеличивается...**
Извините, но Александрийская библиотека уже переполнена...
...так давайте сожжем ее и начнем все заново!
...и **увеличивается...**
У нас есть Google!
...и **увеличивается...**
Теперь у нас есть не просто Google...
...у нас есть Google в квадрате!
...но цель, которую ставит себе статистика, **остается прежней.**
Взглянув на **случайную выборку...**
...мы сможем **выдвинуть предположения о генеральной совокупности,** которую она представляет.

Глава 3

РАНЖИРОВАНИЕ




Мы готовы потратить время и силы на случайную выборку...
...только когда нам любопытно узнать что-нибудь о генеральной совокупности, которую она представляет.


Итак...
Что же вы хотите узнать о нас?
Иногда нам интересно узнать что-нибудь об особенностях опрашиваемых или составить классификацию...
...и иногда хочется скорее сформулировать вопросы, на которые мы можем ответить с помощью полученных цифр.

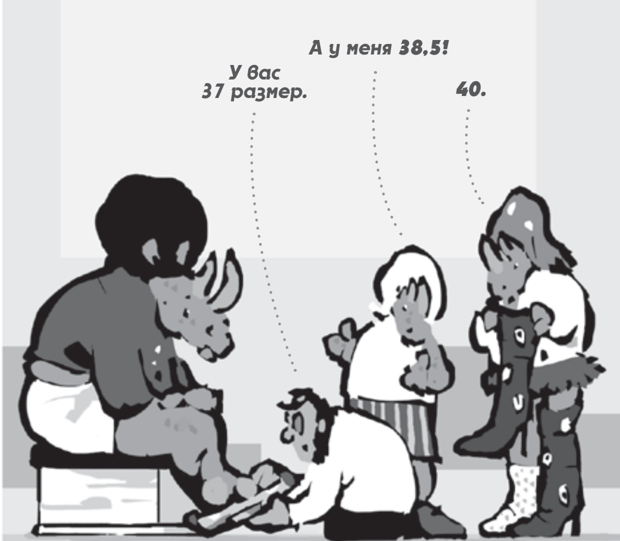

Вы пользуетесь дезодорантом?
Где вы родились?
Что вы любите есть?


Сколько вы спите?
Какова толщина вашей шкуры?
Как часто вы принимаете ванну?









...и эти два вида данных нельзя смешивать.



Мы собираем **категорийные данные**...

...когда изучаем то, что можно **описать только словами**...



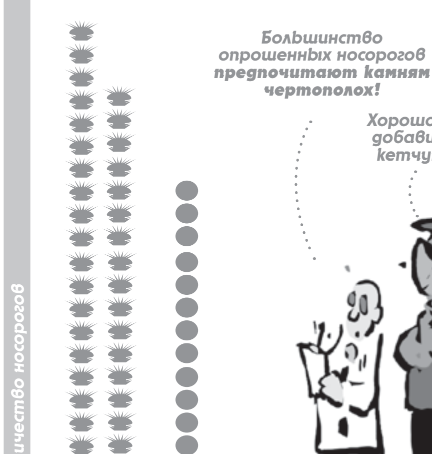




...или когда можно получить **утвердительный/отрицательный ответ.**



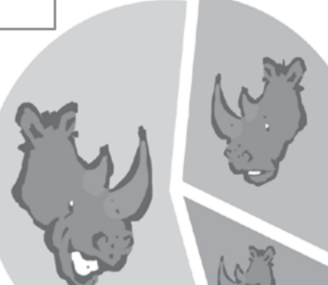
Собрав категорийные данные, мы можем сложить их **стопочкой**...





...**или разделить на кусочки**...

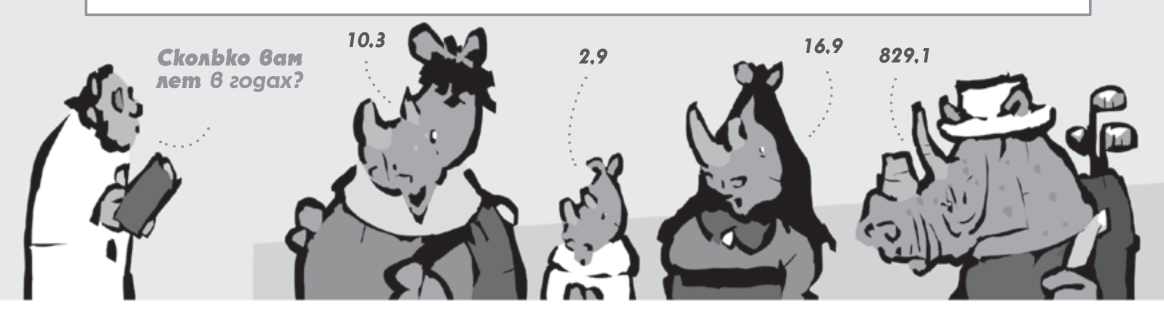


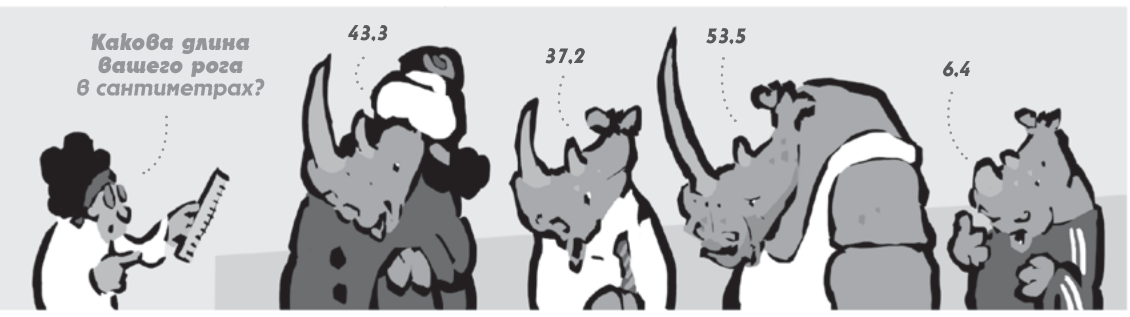


...чтобы можно было получить представление о **соотношении** в нашей выборке.

Мы собираем числовые данные...

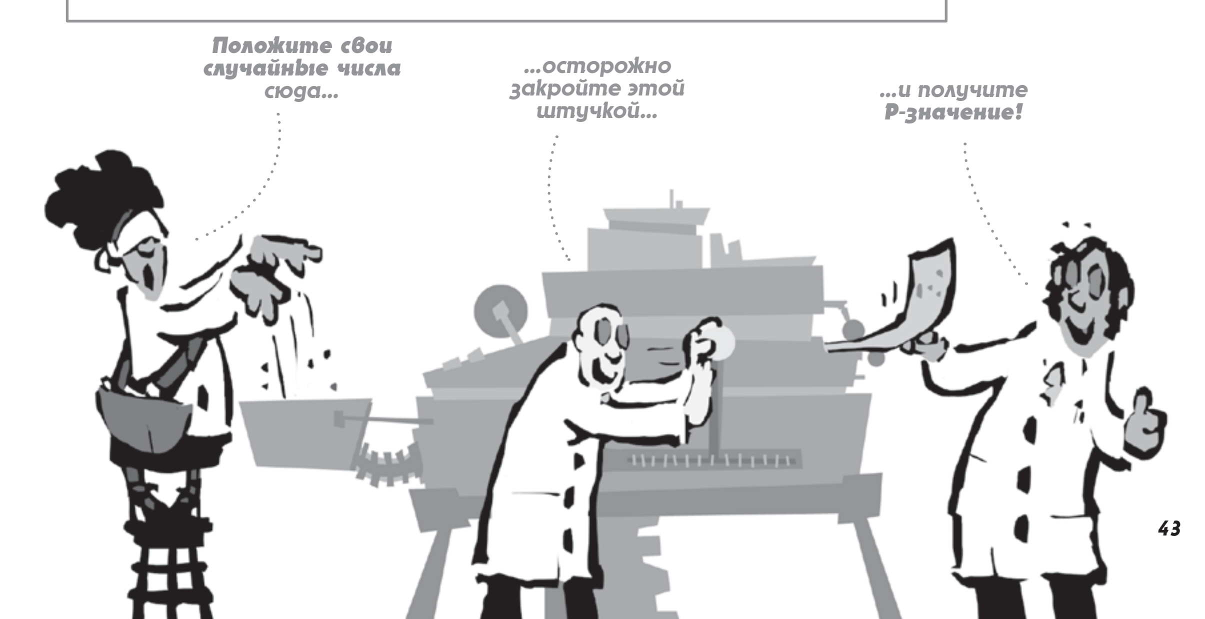
...когда изучаем параметры, которые можно **сравнить, используя числа.**







Как мы увидим во второй части книги, благодаря всем этим показателям числовые данные оказываются в целом гораздо **более полезными.**



Главное различие между двумя этими видами данных...



...заключается в том, что мы **не можем подсчитать категорийные данные...**



...но **можем подсчитать числовые!**



Этот факт превращает в глазах статистиков числовые данные **в нечто захватывающее...**

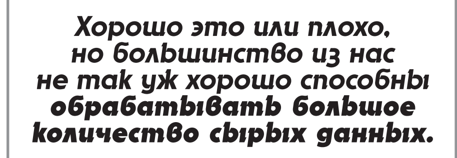


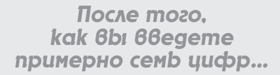


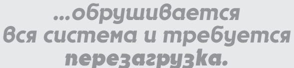
...и кажется чем-то страшным **обычным людям.**

















Поэтому **первое**, что мы делаем, собрав большое количество числовых данных...

Вот 50 случайно отобранных носорогов...

...измерим обхват талии каждого.

У меня талия 290 см.

У меня 333.

Эй, может, хватит налегать на чертополох?

309 314 325 312 309 291 312 319 301 343 309 306 297 341 301 324 341 303 298 291 364 306 308 318 294 317 324 265 305 311 323 293 329 324 295 297 292 313 313 314 309 319 328 308 301 308 291 288

...мы рисуем картинку, где отображаем их все.



Самое простое отображение числовых данных называется *гистограмма.*



Чтобы получить **гистограмму** нашей выборки...

...нарисуем **числовую ось.**



Это те числа, которые расположены на горизонтальной оси между **самыми маленькими...**

...и самыми **большими** объемами, которые мы замерили.



364

Обхват талии (в см) 270 280 290 300 310 320 330 340 350 360 370

Сверху указываем **наши данные...**

...показатель за показателем.

Гистограмма похожа на **огромную башню** из ящиков.

Каждый носорог соответствует **одному ящику.**

Талия у этой дамы-носорога **343** см...

Значит, ей место **вот здесь...**

Количество носорогов 4 5 6 7 8 9 10

Обхват талии (в см) 270 280 290 300 310 320 330 340 350 360 370

Другой вариант визуализации числовых данных представляет собой **коробчатый график/боксплот.**



Чтобы создать боксплот для нашей выборки...

...сначала снова рисуем **числовую ось...**







Обхват талии (в см) 270 280 290 300 310 320 330 340 350 360 370

...но на этот раз помещаем **промежуточные 50%** нашей выборки **в один большой ящик.**



Благодаря этому ящику мы можем понять, где сосредоточена **основная часть данных.**



И с помощью этих планок определяем **минимальные...**

...средние...

...и **максимальные** индивидуальные значения.





Обхват талии (в см) 270 280 290 300 310 320 330 340 350 360 370

*Как правило, **гистограммы** составляют, когда нужно увидеть **полную картину** на основе всех наших данных...*

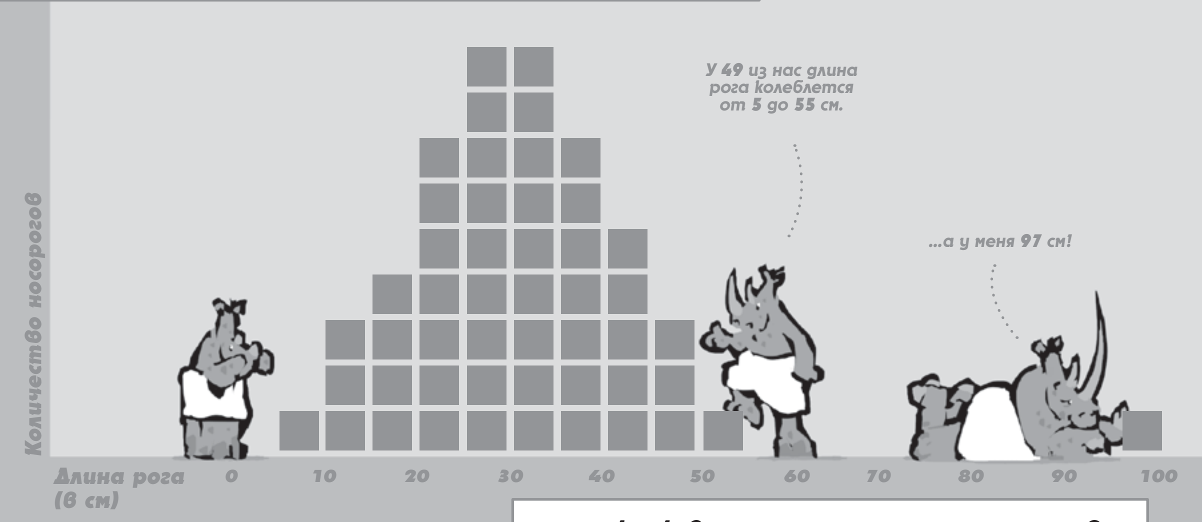
...и выверенных деталей.

Это напоминает **горный хребет.**

Мы можем использовать его, чтобы исследовать **вершины...**

...и **долины.**

*Вот, например, гистограмма, отображающая **длину рога...***



*...показывает, что **один из носорогов намного носорожистей остальных.***

С другой стороны, боксплоты могут быть особенно полезны, если необходимо сделать поверхностный обзор данных...



Боксплот — это компактная версия гистограммы.

Это все равно что смотреть на наши данные из космоса.

...или если мы хотим сравнить разные выборки или группы.

Сравнивая эту выборку...

...с той...

...мы видим, что у нижней в целом более крупные величины.

Благодаря боксплотам — «ящикам с усами» — мы быстро получаем представление о том, как данные собираются воедино...

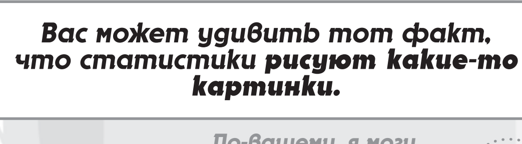
В этой выборке данные группируются гораздо плотнее...

...чем в этой

...и понимаем, к каким выводам они нас приведут.

Странно, большая часть этих данных сдвинута влево...

...а этот огромный кусок — вправо.

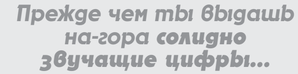

Вас может удивить тот факт, что статистики **рисуют какие-то картинки.**
По-вашему, я могу сделать доверительное суждение о генеральной совокупности...
...опираясь только вот на эту мазню?
Всему, что вам действительно нужно знать, вы научились в детском саду.


Все дело в том, что **первое,** что мы всегда должны делать с собранными данными, это просматривать их.


Вы **удивитесь,** как часто люди об этом забывают.
Потому что, хотя нас могут привлекать **более изощренные математические инструменты...**




Эй, приятель, не интересует **непараметрический иерархический алгоритм Байеса?**
...именно простые картинки будут фокусировать наше внимание на той информации, которую **на самом деле несут собранные данные.**


Прежде чем ты выдашь на-гора **солидно звучащие цифры...**


...нарисуй картинку!


Одна гистограмма стоит тысячи P-значений.

Глава 4
ДЕТЕКТИВНАЯ РАБОТА





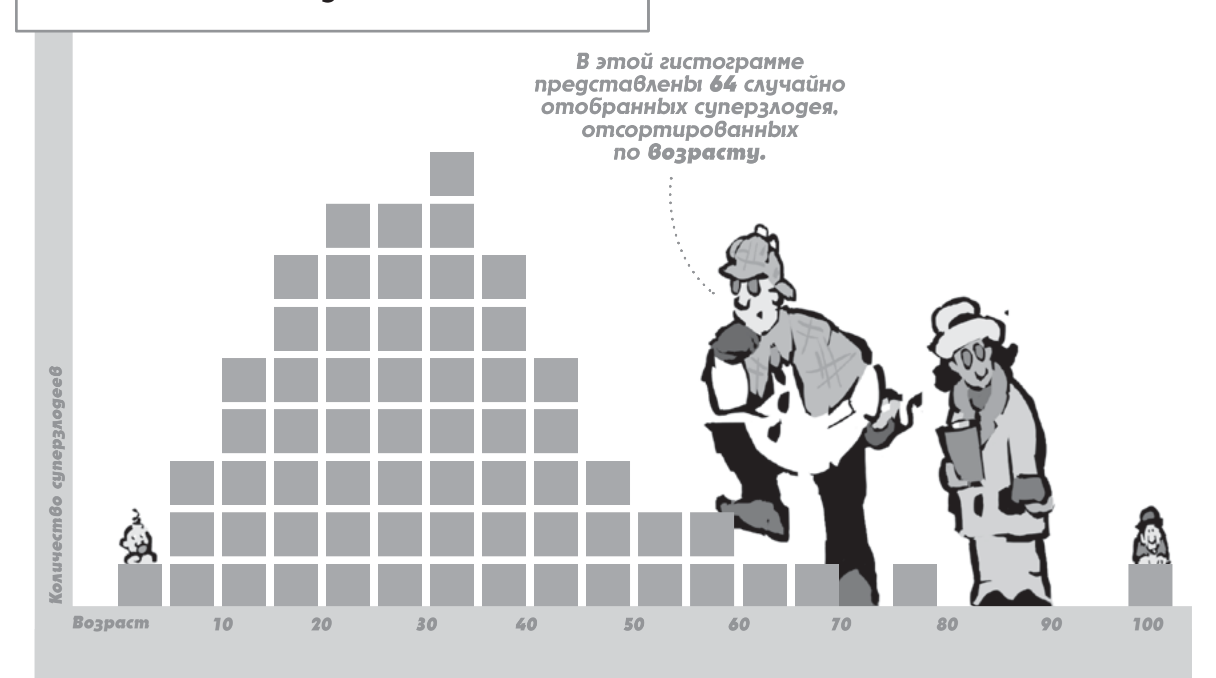
Наша главная цель — сбор **улик** по **одной случайной выборке...**

...и восстановление на их основе **истории генеральной совокупности.**



Но **первым делом** нам придется научиться выполнять самую **простую детективную работу.**

Когда мы только приступаем к анализу любых данных...



...мы всегда обращаем внимание на четыре основные характеристики...









...и мы посвятим эту главу их изучению.



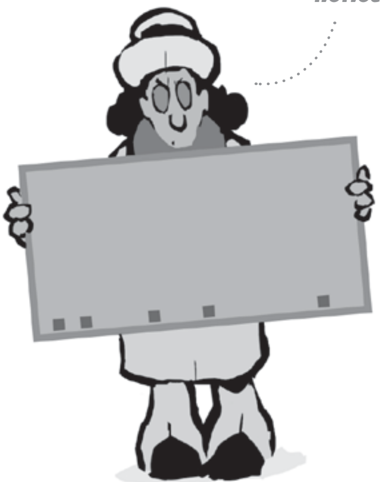
ОБЪЕМ ВЫБОРКИ





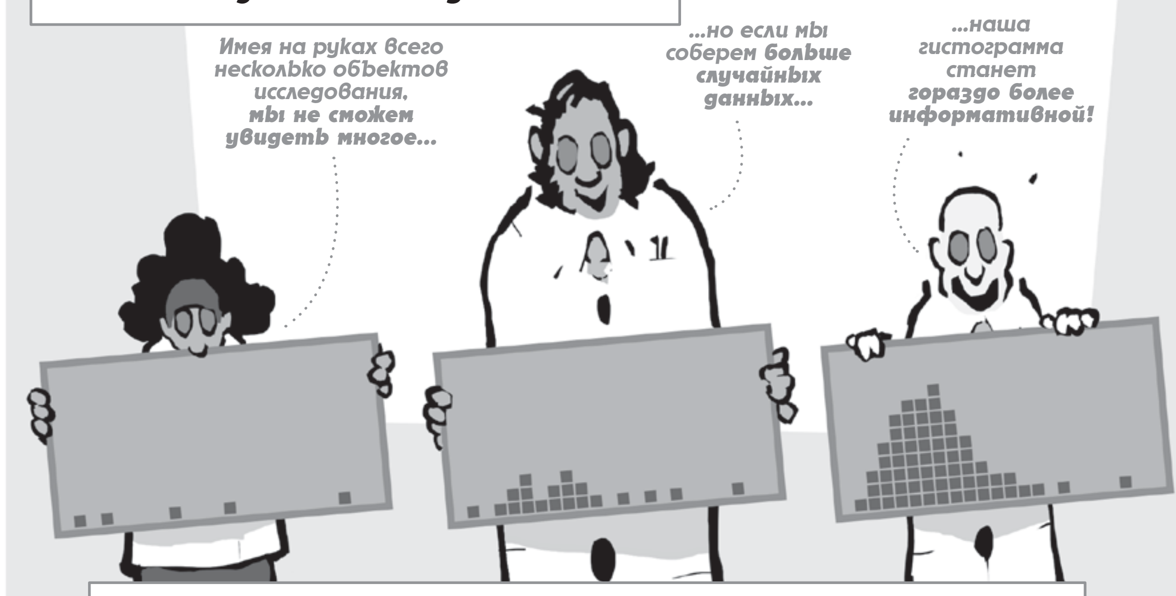




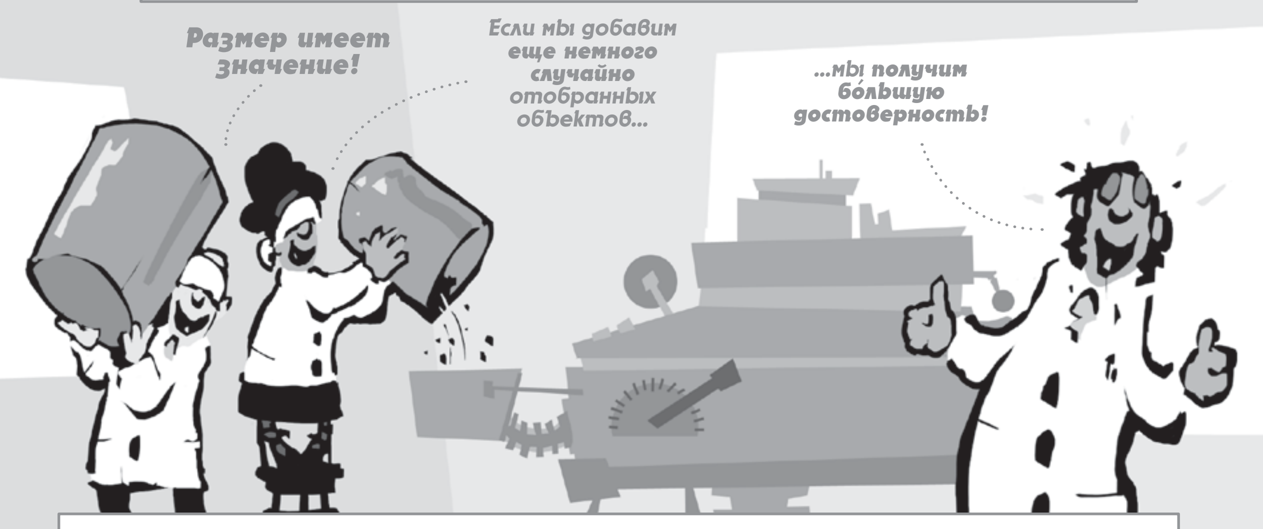


* См. стр. 214.

Как правило, выборка большего размера оказывается полезнее!



Как мы узнаем чуть позже, размер выборки напрямую связан с уровнем достоверности, с которой мы можем судить о генеральной совокупности.



К сожалению, на практике объем выборки всегда чем-нибудь ограничен.





ФОРМА







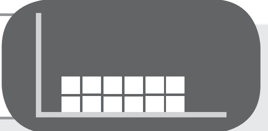
Момент, когда кто-то понимает, какая форма у выборки, может быть весьма захватывающим...



...потому что какой бы ни была ваша гистограмма, она всегда имеет такую форму по какой-то причине.

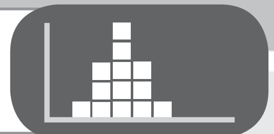


Например, мы называем распределение равномерным, если все исходы одинаково вероятны.



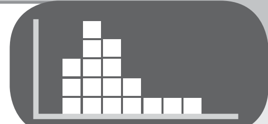


Мы называем распределение данных нормальным, когда есть нечто превалирующее, что заставляет факты группироваться вокруг одного конкретного значения.





Мы называем распределение смещенным, когда по какой-либо причине в одной части находится больше данных, чем в другой.





РАСПОЛОЖЕНИЕ



Под *расположением* понимают место скопления *наибольшего количества данных* на оси.

*Данные могут группироваться вокруг отрицательных **значений**...*

Регулярное употребление моего **волшебного** запатентованного тоника...

...превратит вас в коротышку!

Изменение роста (в см)

-20 -10 0

*...или **маленьких значений**...*

У каждого пирата от **нуля** до **двух** глаз.

К бою!

Количество глаз

0 1 2

*...или **по-настоящему больших значений**.*

Ого, а звезды-то в нашей галактике **старые**!

Возраст в годах

2×10^9 4×10^9 6×10^9 8×10^9

На практике статистиков часто интересует сравнение расположения разных множеств данных.

Эльфов мы, огры, кидаем...

...дальше, чем гномов.

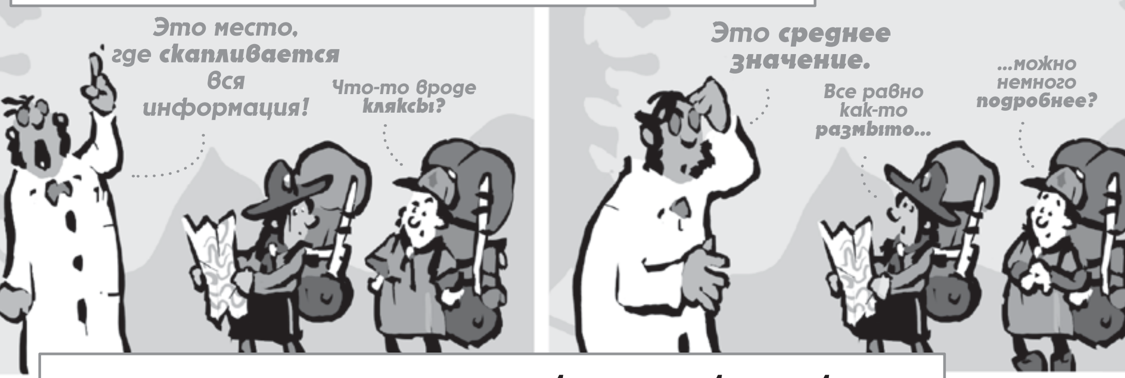
гномы

эльфы

Расстояние в метрах

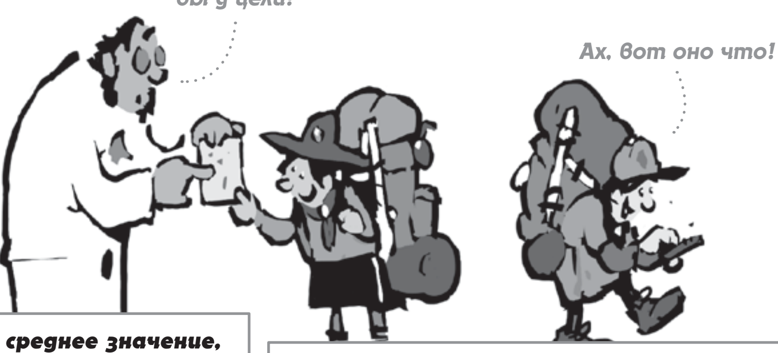
5 10 15 20

Дать слову «расположение» словесное определение может оказаться делом нетривиальным...



...поэтому часто для описания информации мы используем одно значение — среднее*.

* Оно еще называется «среднее арифметическое». Чтобы узнать, как оно высчитывается, откройте стр. 214.



Чтобы подсчитать среднее значение, мы просто складываем все данные...



...и потом делим их на количество данных в совокупности.

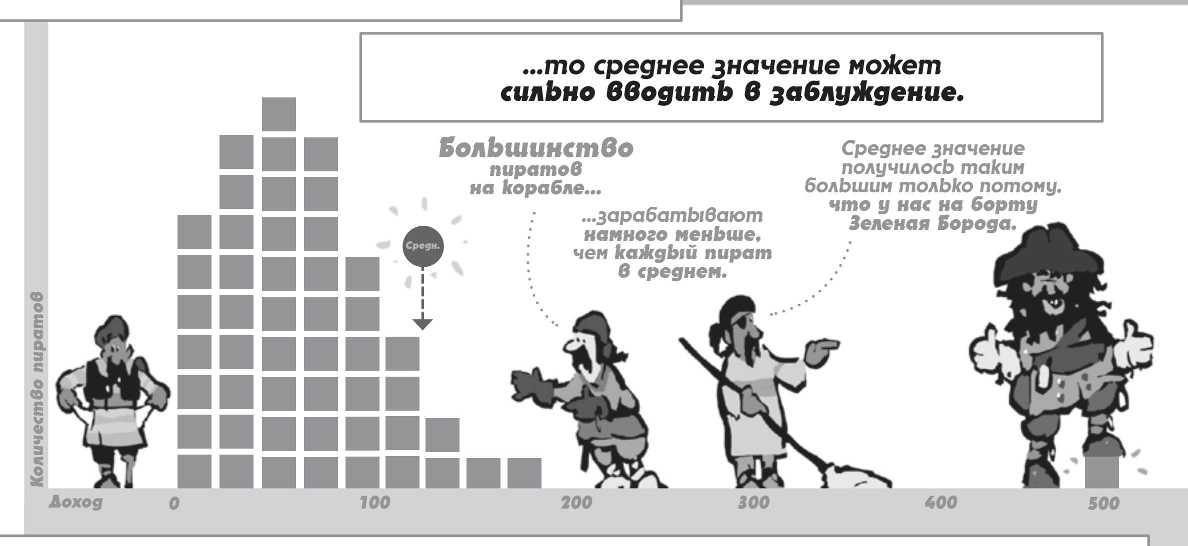


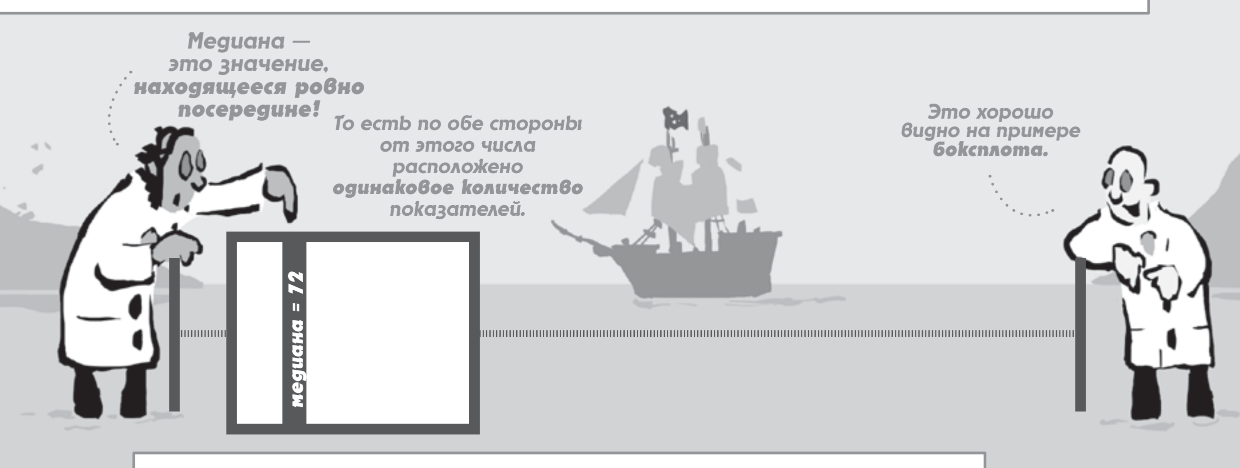
Но даже притом, что среднее значение информативно и точно как средство измерения расположения, оно не идеально.






К сожалению, **среднее значение может быть обманчиво.**
Но тот факт, что наш средний доход составляет 120 дублонов...
...не означает, что большинство из нас настоящие богачи!
Например, если распределение **смещено**...


...то среднее значение может **сильно вводить в заблуждение.**
Большинство пиратов на корабле...
...зарабатывают намного меньше, чем каждый пират в среднем.
Среднее значение получилось таким большим только потому, что у нас на борту Зеленая Борода.
Средн.
Количество пиратов
Доход
0
100
200
300
400
500


В случае смещенного распределения наиболее показательной будет **медиана**...
Медиана — это значение, находящееся ровно посередине!
То есть по обе стороны от этого числа расположено одинаковое количество показателей.
Это хорошо видно на примере боксплота.
медиана = 72


...потому что благодаря этому мы лучше понимаем **«типичное» значение.**
Я — **медиана.**
Я зарабатываю 72 дублона в год.
А я представляю собой **среднее значение.**
Я зарабатываю 120!
Я зарабатываю 500... Йо-хо-хо!
Именно из-за меня показатели этих двоих **так сильно** разнятся.
Доход
0
100
200
300
400
500

Поэтому, когда некоторые с важным видом бросаются средними значениями...



...важно помнить, что эти показатели отображают всего лишь один факт из всего множества данных.





Это одна из причин, по которой никогда нельзя рассматривать расположение множества данных...

...не принимая во внимание еще и форму...



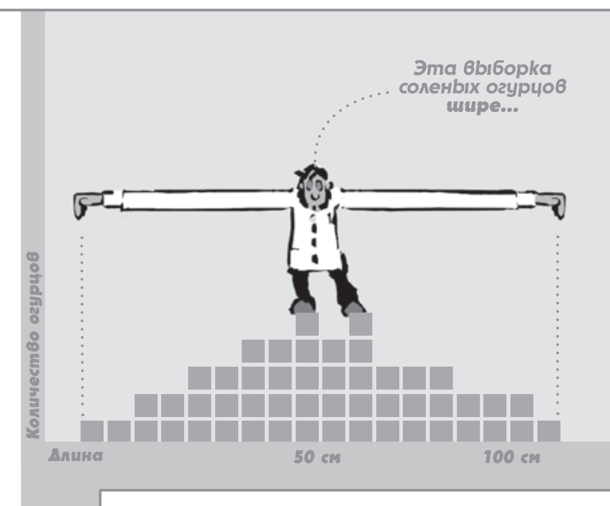
...и размах вариаций, о чем мы поговорим дальше.

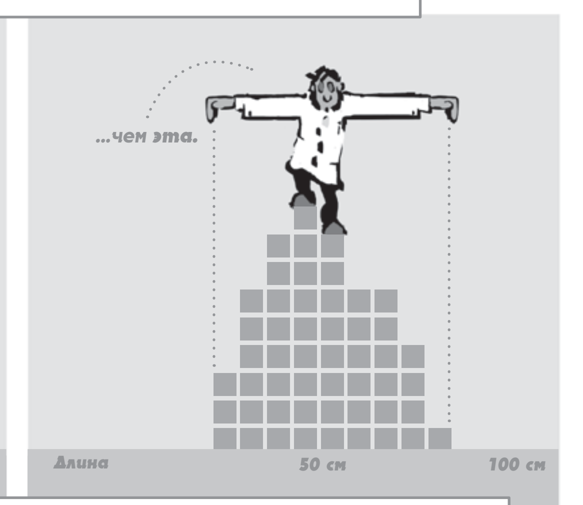


РАЗМАХ ВАРИАЦИЙ



Размах вариаций — это показатель распространенности данных…





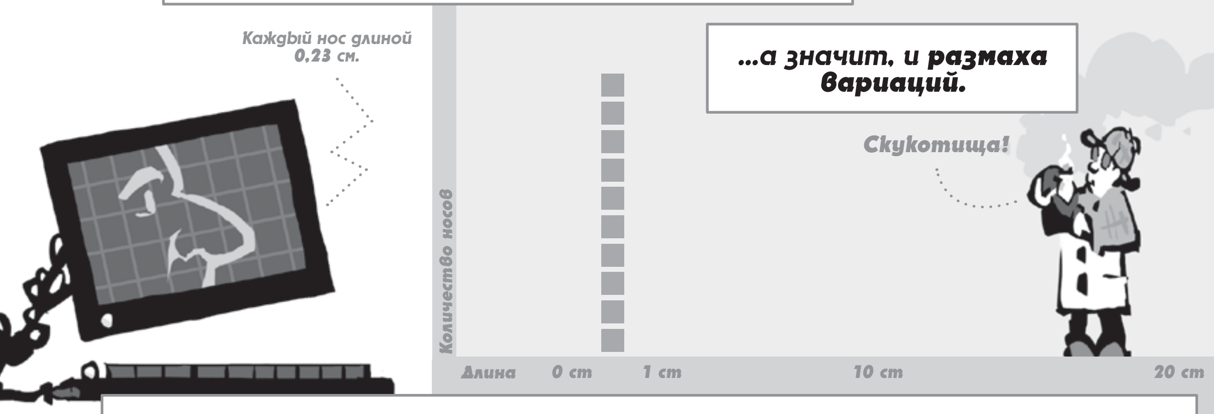
…но это также и мера разнообразия вариантов.





Например, если мы возьмем **выборку из 10** носов, копированных на компьютере...

...мы не увидим **никакого разнообразия**...



Но если мы рассмотрим **выборку из 10 носов, нарисованных от руки...**



...то увидим **большое разнообразие...**





Надежный способ измерить **размах вариаций — взять весь диапазон...**

...который представляет собой разницу между **максимальным** *и* **минимальным** *показателями...*

...и поделить его на **четыре части.**

В каждой части будет **одинаковое число данных.**

В этом случае **16.**

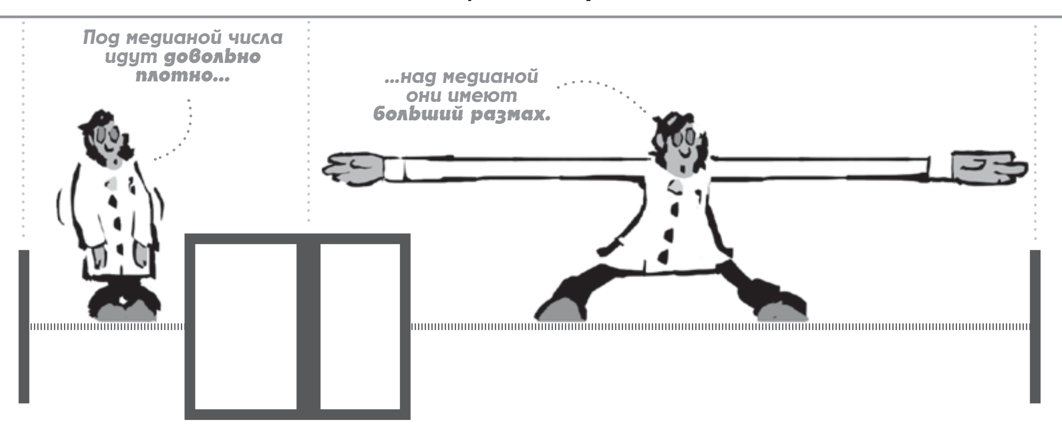
Эти два «куска» в середине называются **«межквартильным размахом».**

Итоговое количество злодеев

Возраст 10 20 30 40 50 60 70 80 90

Размах вариаций

Это дает нам представление о разнообразии **в рамках каждой отдельной части** *общей выборки...*



...и это особенно важно, когда мы исследуем **смещенные данные.**



Наиболее распространенная мера размаха вариаций — **стандартное отклонение (СО)***.



* Чтобы научиться высчитывать его, загляните на стр. 215.

В этой главе мы узнали о четырех важных характеристиках, которые изучаются в любой выборке...



...вскоре мы начнем охоту за этими самыми характеристиками в генеральной совокупности.





Но сначала давайте применим на деле то, чему мы уже научились, чтобы разрешить спор!

Глава 5

СТРАШНЫЕ ОШИБКИ





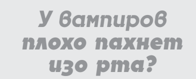








Некоторые вопросы довольно **просты...**



...и ответить на них можно, просто посмотрев **на один набор выборочных данных.**







Но другие вопросы кажутся **неоднозначными...**







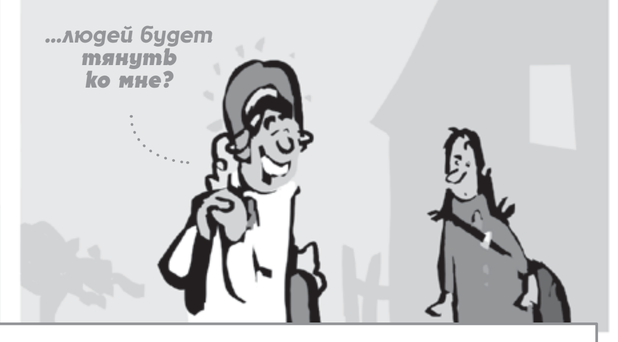
Более сложные статистические проблемы зачастую подразумевают изучение взаимосвязей...











Мы тратим много времени на то, чтобы определить, как сильно одна переменная **влияет на другую...**

...но помните, статистика не может быть абсолютным доказательством **ни одного** из наших выводов.






В этой главе мы собираемся исследовать взаимосвязи двух разных переменных...*
Если бы я был женщиной...
...был бы я более проворным наездником?
...и разрешить-таки спор.
В былые времена только викинги мужского пола объезжали драконов.


Йууухуууу!
Но в последнее время на них стали летать и женщины-викинги...


Йееехуууу!
...и они убеждены, что летают быстрее!

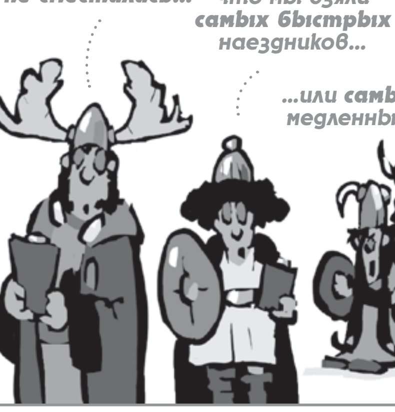

Это так и есть, даже не спорь, ты, шовинистская свинья!

* См. стр. 215.


Чтобы понять, оказывает ли пол...
Итак, наша первая переменная.


...значительное влияние на скорость...
И наша вторая переменная.
...судьи-викинги собрали кое-какие данные.


Они сделали выборку из 50 случайно отобранных наездников...
Победа за нами!

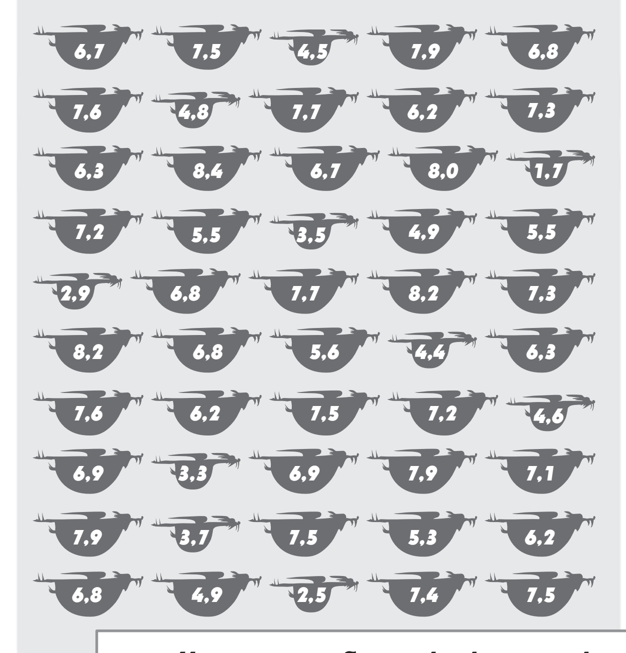

Мы выбрали их произвольно, чтобы результаты не сместились...
...из-за того что мы взяли самых быстрых наездников...
...или самых медленных.


...и 50 наездниц...
Да вы проиграете!


...и засекли время, за которое они преодолеют километр.
Пусть победят самые быстрые наездники!

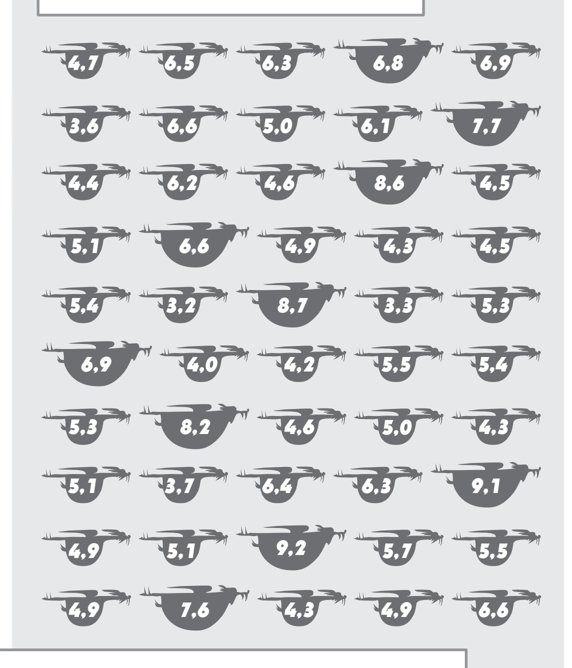


Вот результаты 50 выбранных наугад мужчин-наездников...





...и 50 случайно отобранных наездниц.



Из этого набора сырых данных мы легко можем высчитать два средних значения...





...и сравнить их.





Благодаря этому сравнению конфликт, кажется, исчерпан.



Но мы пока разобрали только **одну составляющую общей картины.**





Чтобы получить более точное представление о данных, нам **необходимо нарисовать картинки.**

Подойдите поближе и давайте посмотрим, **о чем нам говорят эти цифры.**

Это было вполне предсказуемо, что картинки получатся разные...

Обе группы смещены?

...если мы сравним боксплот с показателями мужчин...

Так вышло, что в целом мы преодолели расстояние медленнее...

*...зато у нас **больше вариаций** по скорости, которые видны **на более быстрой стороне!***

медиана

Секунды 1 2 3 4 5 6 7 8 9 10

...с аналогичным исследованием женских показателей.

*Мы затратили **меньше времени на прохождение дистанции...***

*...но у нас получилось **больше вариаций** на **более медленной стороне!***

медиана

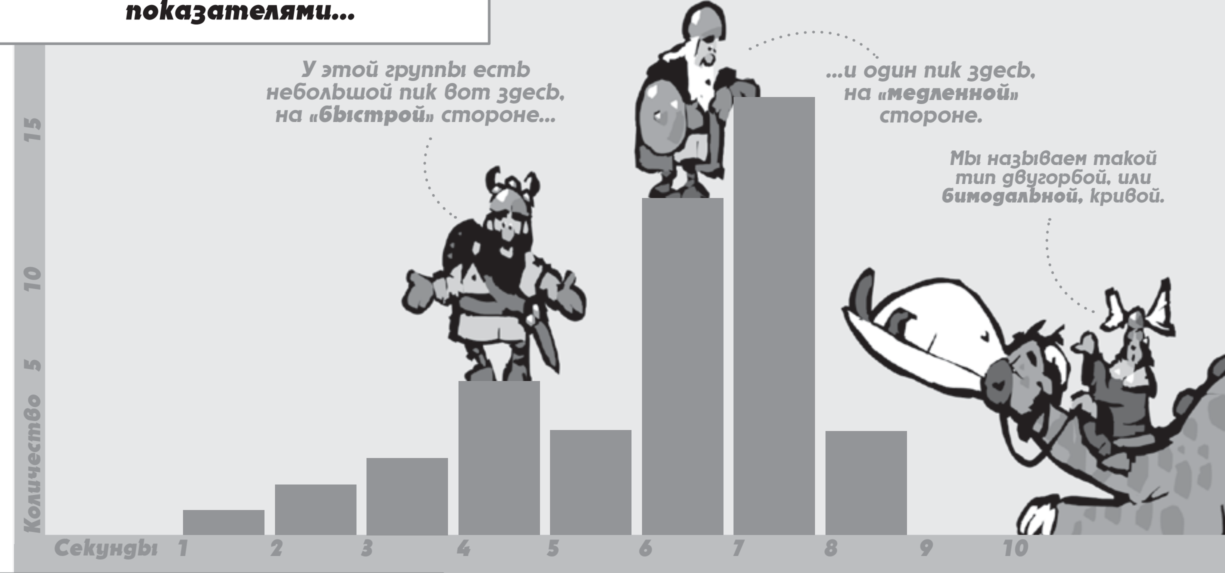
Секунды 1 2 3 4 5 6 7 8 9 10



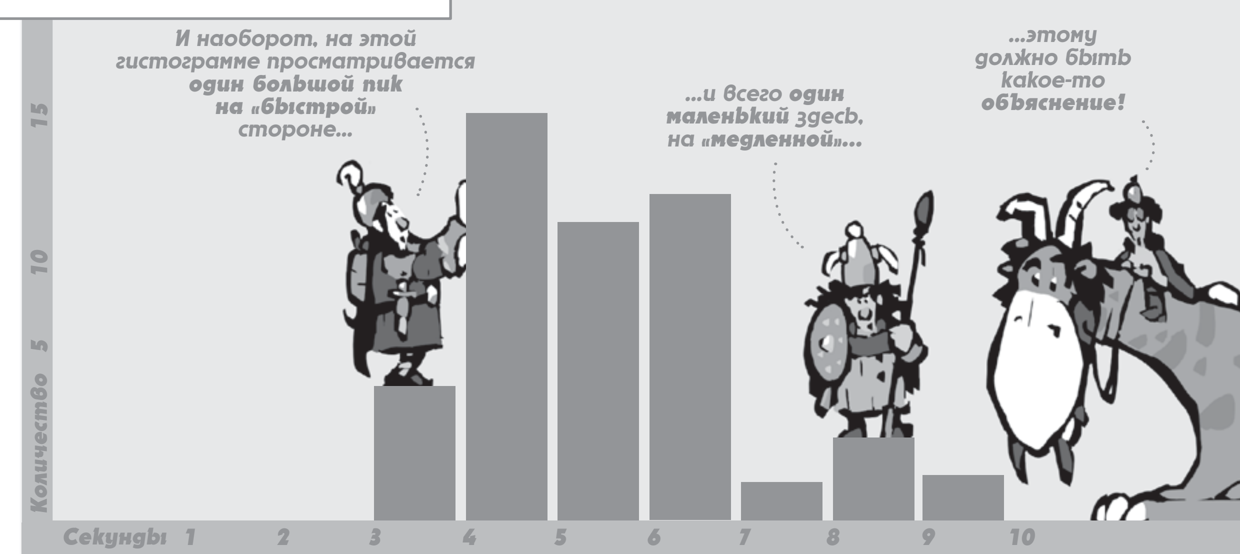
И мистика только усиливается...







...и на женские показатели.



Это только доказывает, что соотношение двух наших переменных может на деле оказаться не таким простым, как мы думали!

Если тот **факт, что ты женщина,**

...заставляет тебя **летать быстрее...**

...тогда почему на обеих гистограммах **есть смещения и присутствует мистическая двугорбость?**

Помните, какими бы ни получились гистограммы по вашим данным, на то всегда есть **причина.**



Теперь основная задача в том, чтобы понять, почему данные выглядят именно так...



...мы можем выяснить это, поискав другие переменные, которые могут оказывать влияние.







Выясняется, что пока мы концентрируемся на половой принадлежности участников и скорости...

...мы совершенно забываем о драконах!





Дело в том, что драконы бывают двух видов...





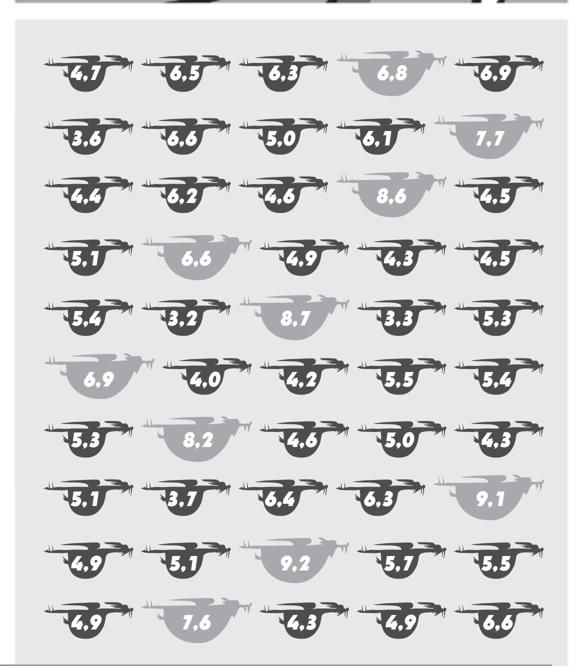
*...и мужчины-наездники, как правило, предпочитают драконов покрупнее, которые оказываются **менее расторопными...***



*...в то время как дамы отдают предпочтение драконам поменьше, но **пошустрее!***

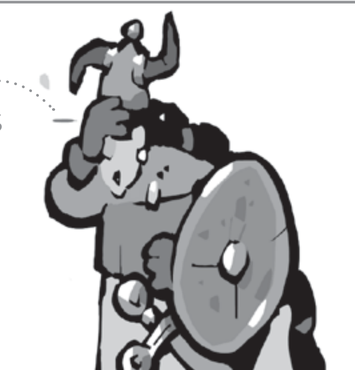






Получается, неудивительно, что наездницы в целом оказались быстрее.





Если мы примем во внимание тот факт, что **наездники разного пола предпочитают неодинаковых драконов...**

Ну да, я люблю **драконов побольше**, ты что-то имеешь против?

...то при подсчете среднего времени **для обоих типов наездников и драконов...**

...мы получим весьма **неожиданные результаты.**



	миниатюрные драконы	крупные драконы
наездники	3,6	6,9
наездницы	5,1	7,9



Выходит, что наше **первое заключение...**

...оказалось не просто **обманчивым...**





...а в корне неверным!









...мы совсем забыли о возможной третьей переменной, которая все это время была где-то поблизости...







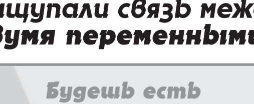
К сожалению, скрытые переменные могут внести неразбериху в любой статистический анализ...



Мораль этой истории заключается в том, что....

*Будьте бдительны и помните про **скрытые переменные!***

...всякий раз, когда нам кажется,** что мы нащупали связь между **двумя переменными...





...может статься, что есть и другие факторы, оказывающие влияние на наши заключения...











*...и если мы их не найдем, мы рискуем поверить в то, что **на деле не будет правдой!***

Глава 6

От выборки к генеральной совокупности





Пока что мы говорили в основном о **выборках.**

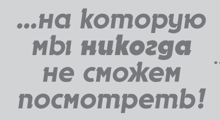
Вот у нас тут есть 50 рыбешек, выбранных наугад и распределенных по весу!

Но помните, наша **конечная цель** — использование **выборки...**

...на которую **мы можем посмотреть!**



...для получения объективных выводов о генеральной совокупности.



...на которую мы **никогда** не сможем посмотреть!







И это создает *проблему*:





Во второй части нашей книги мы поставим этот вопрос ребром...

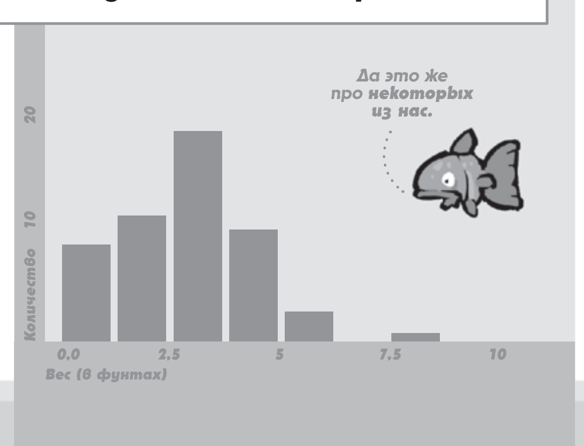


...но прежде чем мы начнем, давайте проясним некоторые *ключевые термины*, которые будем использовать.

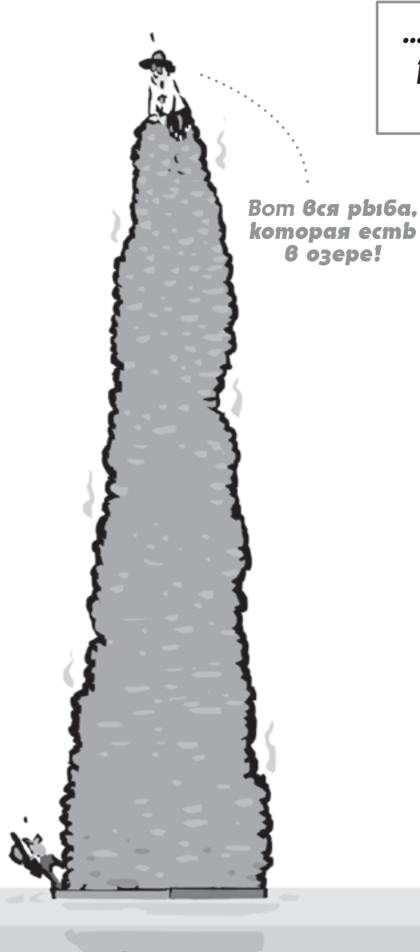
Мы уже знаем, что упорядоченные данные нашей выборки в виде графика...

...называются гистограммой.





Но если бы нам удалось собрать вместе данные обо всей генеральной совокупности...



...мы бы назвали получившийся результат распределением генеральной совокупности*.

600
500
400
300
200
100
Количество
0.0
2.5
5
7.5
10
Вес (в фунтах)

Эта штука про всех нас.

Запомните, в реальности вам никогда не удастся увидеть целиком все совокупное распределение...

...если бы это было в ваших силах, вы бы спокойно обошлись без статистики.





* См. стр. 216.

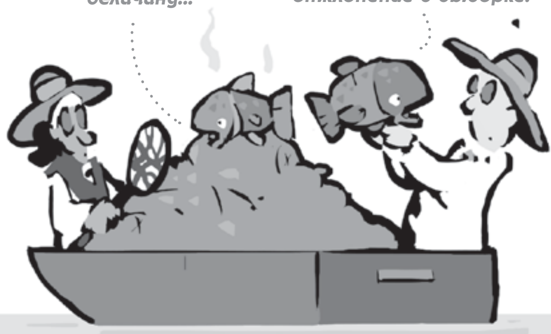
Мы уже знаем, что у выборки на гистограмме есть определенные важные показатели...



...а тут вдруг оказывается, что у совокупного распределения тоже есть эти показатели.

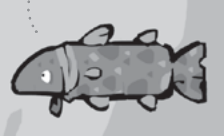


Чтобы различить их между собой, мы называем показатели в выборке «статистическими величинами»...



...а показатели в совокупности — «параметрами»*.



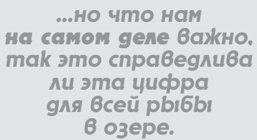



* См. стр. 216.

Иными словами, единственное, что заставляет нас отправляться **собирать статистические данные...**

...это наше **любопытство:** нам интересно, какими будут **параметры.**









Статистические данные — это то, что мы, **собственно, подсчитываем** и **о чем можем судить с всей определенностью...**

...а **параметры** — это то, что мы бы **хотели знать,** но о чем можем **только строить предположения.**









Пусть мы и **никогда** не сможем **посмотреть на параметры своими глазами...**

...но, к счастью, у нас есть **статистические данные,** чтобы **определить параметры.**



Прихвати с собой статистические данные...

...мы отправляемся на **ответственное задание!**




На самом деле статистические данные помогают найти **самые разные виды параметров.**
Стандартное отклонение.
Пропорции.
Медианы.
Назовем их!
Мы подробно остановимся **на каждом отдельно.**


Мы будем учиться использовать статистические данные, которые находим в **случайной выборке...**
...чтобы определить средние значения в совокупности, которую она представляет.
Объем выборки?
— Есть!
Среднее значение выборки?
— Есть!
Стандартное отклонение в выборке?
— Есть!
Итак, мы готовы, **пойдем искать параметры!**
Они должны быть где-то тут!

Как нам уже известно, мы никогда не сможем использовать **статистические данные...**

...чтобы **определить параметры с точностью.**



Часть вторая

ПОИСК ПАРАМЕТРОВ



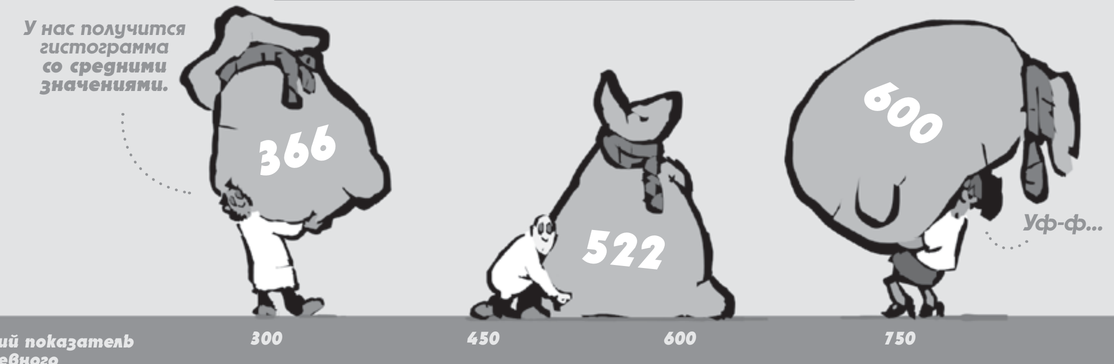
Глава 7

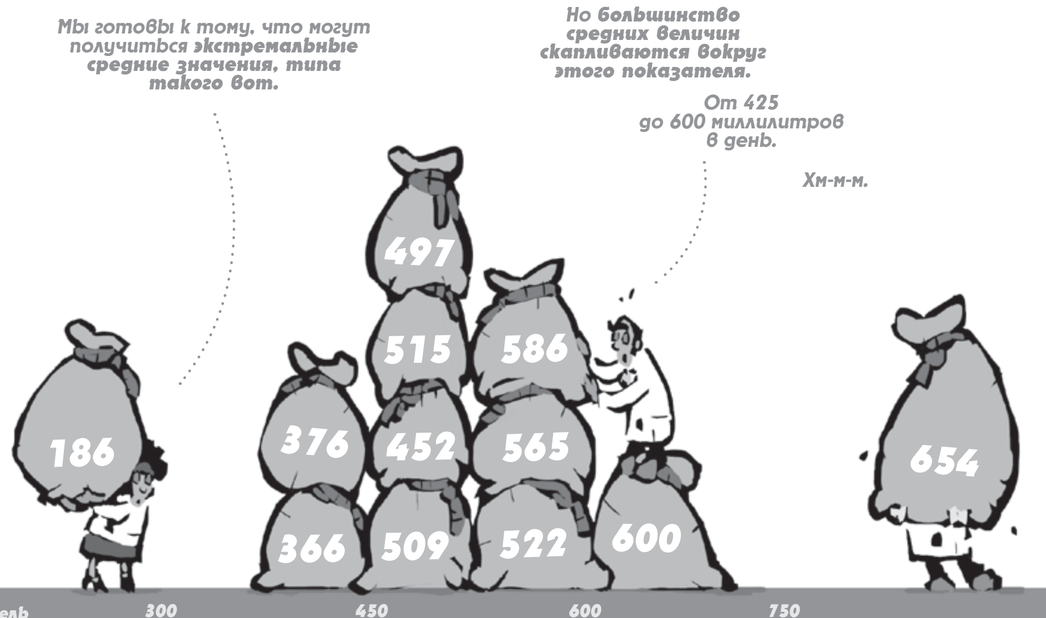
ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА



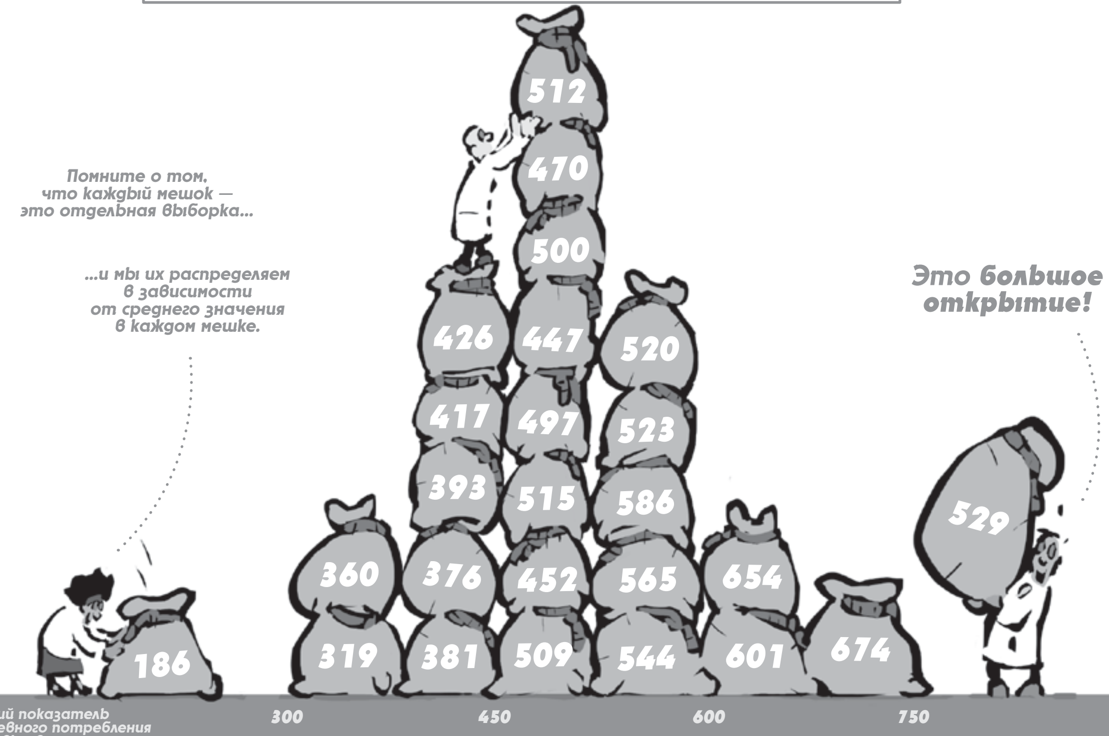

Давайте представим себе, что нам нужно узнать среднее значение в определенной совокупности.
Сколько газировки американец выпивает в день?
Вкусно!
Класс!
А теперь представьте, что мы идем и делаем множество случайных независимых выборок из генеральной совокупности.
У нас есть 50 выбранных наугад американцев.
Тут еще 50 выбранных наугад американцев.
И еще одна группа из 50 выбранных наугад американцев.
В каждой выборке 50 выбранных наугад американцев.
Мы складываем каждую выборку в мешок, чтобы было проще следить за ними.
Эй, чур, не толкаться!


476
Оказывается, если мы высчитаем
среднее значение в каждой выборке...
Например, среднее значение в нашей выборке — 487 миллилитров.
А в нашей — 366 миллилитров.
А тут 522 миллилитра.
186.
452
487
366
522
186


...а потом расставим их по порядку
и разместим одна на другой...
У нас получится гистограмма со средними значениями.
366
522
600
Уф-ф...
Средний показатель ежедневного потребления газировки, в мл
300
450
600
750
...все это множество средних значений в конце концов
сгруппируется!


Мы готовы к тому, что могут получиться экстремальные средние значения, типа такого вот.
Но большинство средних величин скапливаются вокруг этого показателя.
От 425 до 600 миллилитров в день.
Хм-м-м.
497
515
586
186
376
452
565
654
366
509
522
600
Средний показатель ежедневного потребления газировки, в мл
300
450
600
750
И это еще **не все.**







* См. стр. 217, там об этом написано более подробно.

И эта формула подходит для вычисления среднего значения выборки из любой совокупности.

Вот пирамида, составленная из разных видов чешуй драконов...

...отобранных в случайном порядке и распределенных по среднему весу.

А вот случайные выборки лап ящериц...

...отобранных по средней длине.

И неважно, какой формы сама по себе совокупность.

Она может быть такой...

...или такой!

Равномерная форма, смещенная, обычная, ненормальная — да какая разница!

В конце концов, чем больше средних значений вы соберете, тем более нормально-распределенную форму получите.

Куполообразная и симметричная форма!

Очень плавно нисходящая с обеих сторон.

Это самая красивая форма в статистике!

Частота

Средние значения

Формально вот такое огромное множество представляет собой уже вид распределения выборки*.

Так статистические данные выборки были бы распределены...

...если бы мы собрали сто тысяч миллионов образцов.

* Определение см. на стр. 217.

Ну и небольшой приятный бонус:



оказывается, что центральный показатель в огромном множестве средних значений...



...равен центральному показателю генеральной совокупности, которую представляет выборка.

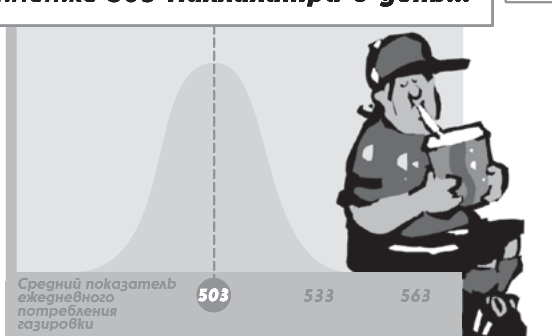
Среднее значение всех средних значений...

равняется среднему значению генеральной совокупности.

Генеральная совокупность может иметь и такую форму...

...мы никогда не можем знать наверняка.

Например, если это множество средних значений в выборке, составленной по количеству газировки, выпиваемой ежедневно, будет центрировано по отметке **503 миллилитра в день...**



...то генеральная совокупность **будет центрироваться по этому же показателю!**

Средний показатель ежедневного потребления газировки 473 503 533 563

Это происходит потому, что огромное множество средних значений **гарантированно будет иметь симметричную форму.**

Нормальное распределение **всегда** симметрично.

В конце концов для **каждого** среднего значения выборки, получаемого с помощью показателя, который **ниже среднего значения генеральной совокупности...**

Конкретно эти 50 случайным образом отобранных **американцев пьют немного газировки.**



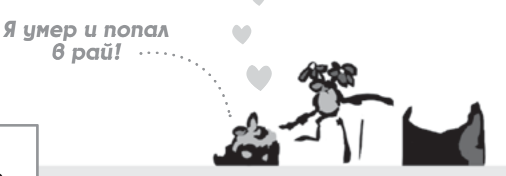
...мы гарантированно получим **другое** среднее значение выборки с помощью показателя, который **выше среднего значения совокупности.**

А эти 50 случайным образом отобранных американцев **пьют очень много газировки.**



А вот и еще один приятный бонус:

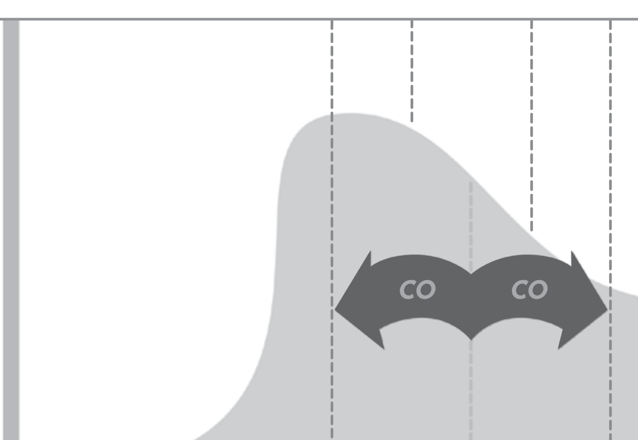




оказывается, что **огромное множество средних значений...**



...как правило, тоже будет у́же, чем генеральная совокупность, которую оно представляет.





А вот насколько у́же, будет зависеть от **размера каждой выборки.**

Если мы увеличим размер выборки, то самое большое множество будет выглядеть скорее **не так...**

Короткий и широкий холмик.

...а вот **так.**





Можно включить интуицию и понять, почему больший размер выборки дает более узкое множество средних значений.



Если в каждой выборке только один американец...

...то размах выборки столбца средних значений будет ровно таким же, как размах выборки генеральной совокупности.



Разница между двумя мешками...

...будет равна разнице между отдельными особями в генеральной совокупности!

Но если уместить в каждой выборке всех американцев из генеральной совокупности...

.... ТОГДА распределение столбца средних значений будет равно НУЛЮ.





В любом случае, математическое соотношение — понятие точное.















**получается, что
огромное множество средних значений
случайной выборки стремится
К НОРМАЛЬНОМУ РАСПРЕДЕЛЕНИЮ!**

Помните,
все выборки
одного и того же
объема.

Они все из одной
и той же генеральной
совокупности.

И их сотни тысяч
миллионов!

Ты выборочное
распределение
моих желаний.



Только
посмотрите
на эти кривые.
Как они красивы!



Количество корзин (75 случайных цветков в каждой корзине)

Цветочные корзины сортированы по среднему размеру

**Все они центрированы по среднему
показателю генеральной совокупности...**

**...но их распределение
у́же, чем у генеральной
совокупности.**

И неважно,
как распределены
выборки...

...и какова
форма...

...или
генеральная
совокупность!



Официально мы называем это открытие ***центральной предельной теоремой (ЦПТ)****.*



* Откройте стр. 217–218, чтобы узнать о ЦПТ подробнее.

За долгие годы статистики выработали формулы, которые объясняют, почему ЦПТ работает.



Она работает только в том случае, *если каждая выборка будет* ***случайной...***

Только по воле ***случая*** *одна выборка отличается от любой другой.*

*...**а также** при условии, что выборка* ***достаточно большая.***

Размер выборки **от 30 и больше** считается достаточным...

...но все зависит от других сложных математических показателей.

* **Откройте стр. 217**, чтобы узнать, как все это называется у математиков.

А вот что ЦПТ представляет собой в математических терминах:*

Мы, конечно, можем ожидать, что огромные множества средних значений выборки окажутся **стандартными**...[1]

...и будут **центрированы по среднему значению генеральной совокупности...**

...со **стандартным отклонением,** равным...

...стандартному отклонению генеральной совокупности, поделенному на квадратный корень объема выборки[2].

Вот это да!

1. Но только если выборки делаются случайным образом и размер каждой достаточно велик (больше 30 или около того).

2. Для любителей математики: обратите внимание, что весь прямоугольник будет у́же, если размер выборки **больше**.



Но вот способ ***попроще,***
как все это запомнить:



...вот в таком прекрасном виде!

Из нескольких следующих глав мы узнаем,
почему это имеет такое значение.




Мы не знаем
всего...
...но это
не означает,
что **мы не знаем
ничего!**

Глава 8

ВЕРОЯТНОСТИ



Из предыдущей главы мы узнали, что **огромное множество средних значений выборки...**

...обычно стремится **к нормальному распределению.**

Количество кувшинов (500 случайным образом отобранных оливок на кувшин)

Кувшины с оливками, расставленные с учетом среднего веса оливок в них

Мы можем заявлять об этом с **уверенностью...**

...при условии, что это были **случайные** выборки...

...и **объем выборки достаточно большой!**

Мы собираемся узнать, **почему же это так важно...**

И что такого **необычного в этой форме?**

...изучив **огромное множество средних значений выборки...**

...**в сарайчике Безумного Билли, где он хранит свои снасти.**

Crazy Billy's
BAIT BARN
Сарайчик Безумного Билли

Привет! **Я Билли!**

Будьте осторожны...

...он сумасшедший!

Безумного Билли так называют, потому что он проводит **сумасшедшее** количество времени, создавая случайные **выборки червей...**



...складывая каждую выборку в консервную банку...

Прежде чем запечатывать банку, я замеряю всех червей....

...и вычисляю среднюю длину червя в каждой банке.

...и составляя одну на другую **сто тысяч миллионов таких банок,** где каждая соответствует своему среднему значению...

...в своем **безразмерном сарае, где он хранит снасти...**

Средняя длина червей в этой банке 4,75 см, значит, ее надо ставить именно сюда.

Количество банок

Средняя длина червей в банке (в см)

3,5 4 4,5 5

...во всяком случае, так он уверяет.





В этой главе мы выясним, что можно узнать об огромном множестве, которым располагает Безумный Билли.





Более подробно мы остановимся на **следующих вопросах:**

если у нас будет доступ только к тому, что внутри сарайчика...

...что же мы сможем сказать о генеральной совокупности червей в болоте?



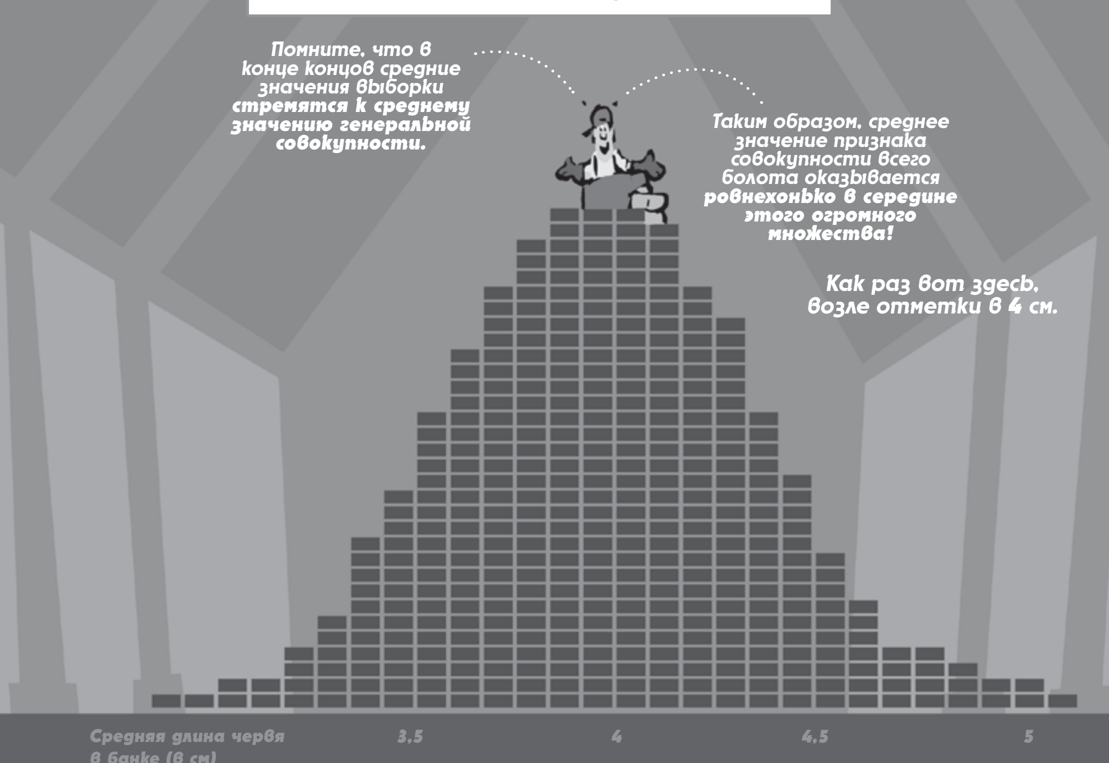
А вот тот же вопрос, но **научными терминами.**






Первый важный вывод, который мы бы сделали, если бы нам удалось хотя бы мельком заглянуть в сарайчик Билли...

...касался бы среднего значения генеральной совокупности.



Иными словами, если бы нам нужно было вычислить среднее значение генеральной совокупности в болоте...

...мы могли бы просто заглянуть в сарайчик — и нашли бы его там!





Но это еще не все...

Другим важным открытием, которое мы могли бы сделать, заглянув в сарайчик Билли...



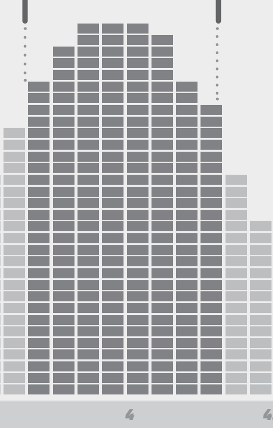
...был бы подсчет вероятностей в отношении генеральной совокупности!



И вот как это работает:

если бы мы могли **подсчитать все консервные банки в огромном множестве, которым располагает Билли...**

...и обнаружить, что у **50% из них** среднее значение колеблется в этих пределах...



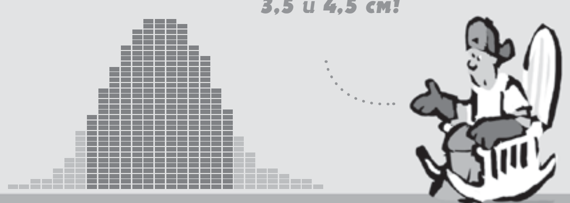

...это бы означало, что, забери мы **одну случайную банку** из **совокупности...**

...с **50%-ной вероятностью** ее среднее значение находилось бы **в тех же пределах!**





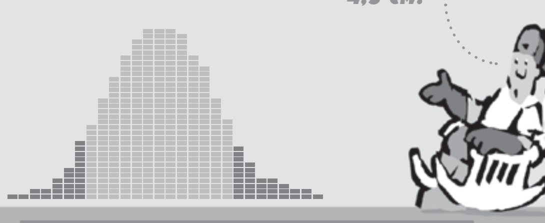
Если бы мы подсчитали все консервные банки и обнаружили, что в **множестве, собранном Билли:**



Это бы означало, что **в генеральной совокупности:**



А если предположить, что в **этом** множестве:



...то можно было бы сделать **такие** выводы о совокупности:



Иными словами, заглянув мельком в сарай...

...мы можем посчитать, каков диапазон средних значений экземпляров, собранных со всего болота!





Есть несколько вещей, о которых нужно помнить при подсчете вероятности*.



* См. стр. 218.

Во-первых, вероятности актуальны только в долгосрочной перспективе...

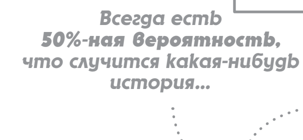
...поэтому они **никогда не скажут ничего достоверно о коротком периоде.**

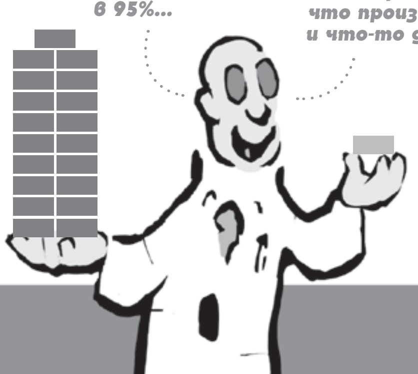


Во-вторых, у каждой вероятности есть обратная сторона...



...потому что вероятности **составляют 100%.**





И наконец, мы, по определению, можем высчитать, с какой вероятностью произойдут события, только если они происходят случайно...

Вероятность, по определению, означает степень возможности наступления определенного события в долгосрочной перспективе.

...вот почему мы **собираем статистические данные только случайным образом.**



Говоря общо, мы можем высчитать вероятность других **случайных событий, например при подбрасывании монетки...**

Вероятность того, что, **подбрасывая монетку,** вы получите решку...

...составляет 50%...



...или броске **игральных костей.**





Но давайте-ка вернемся к **поиску червей случайным образом...**

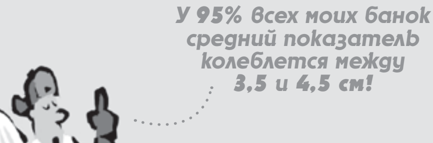


...потому что нам **предстоит узнать еще кое-что важное о сарайчике Билли!**

Оказывается, нам совсем не обязательно пересчитывать все банки в сарайчике Билли...









Оказывается, зная, что все множество банок Билли имеет нормальное распределение...

Это и есть центральная предельная теорема!

Вау!

...мы можем использовать занимательную математику, чтобы вычислить, как банки располагаются внутри него!

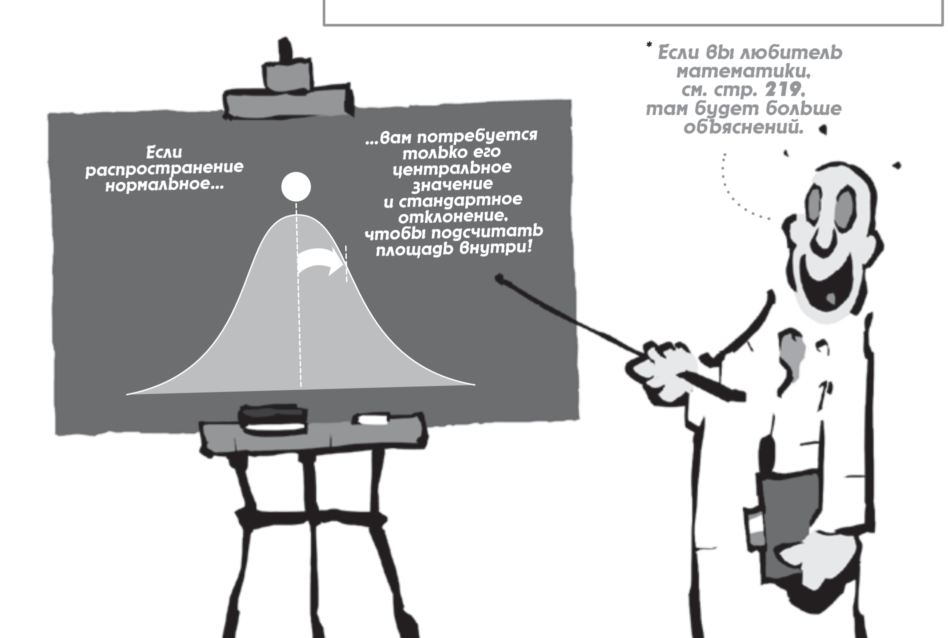
Средняя длина червя в банке, см

3 | 3,5 | 4 | 4,5 | 5



Более того, именно потому что множество распределено нормально...

...нам нужно всего лишь знать его центральное значение и стандартное отклонение, чтобы все высчитать*.



* Если вы любитель математики, см. стр. 219, там будет больше объяснений.

А вот настоящие подсчеты, которые предполагает классическая математика, на самом деле очень сложны.



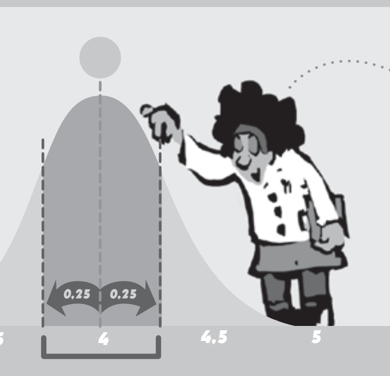
К счастью, есть так называемое *правило большого пальца*, которое действует в случае любого нормального распределения:





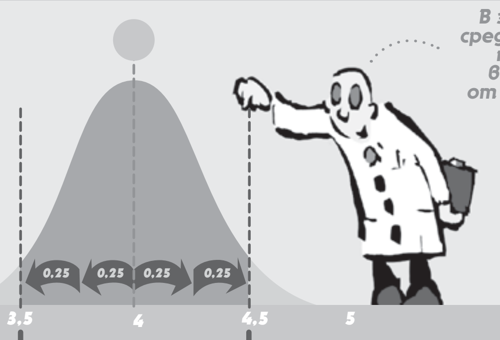
68% всех консервных банок...

...находятся в пределах ***1 стандартного отклонения*** от центра.



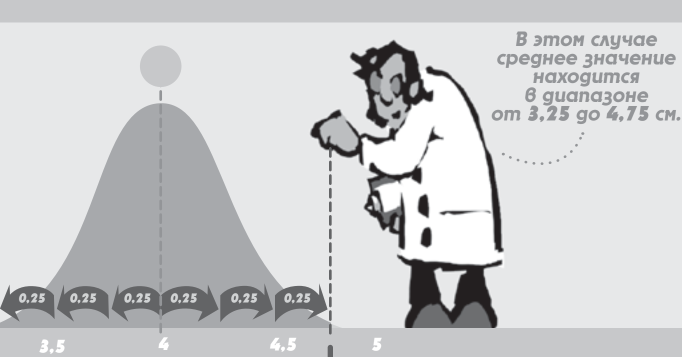
95% всех консервных банок...

...находятся в **2 *стандартных отклонениях*** от центра.



99,7% всех консервных банок...

...находятся в **3 *стандартных отклонениях*** от центра.



Если создается впечатление, что все эти цифры только **сбивают с толку...**



...просто сконцентрируйтесь на **затемненных областях.**

Очевидно, что **внутри** этой затемненной области, представляющей собой часть множества Билли, **намного больше** банок...

...чем **вот здесь.**



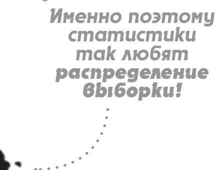


3 3.5 4 4.5 5

Что тут важно помнить, так это то, что затемненные области **внутри распределения выборки Билли...**

...напрямую соотносятся с нашими шансами собрать средние значения **из болота!**







Давайте-ка **резюмируем.**

Первая замечательная вещь, *которую мы узнали о распределении выборки Билли...*

...это то, что оно показывает нам ***среднее значение генеральной совокупности!***





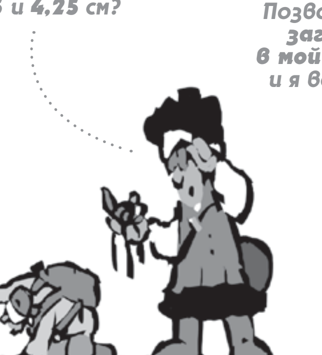
Вторая замечательная вещь, *касающаяся распределения выборки Билли...*

...это то, что мы можем использовать его, чтобы ***высчитывать вероятность для всей генеральной совокупности...***







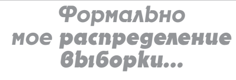




Ясно, что, если бы мы охотились за **средним значением во всей генеральной совокупности...**



...распределение выборки по типу того, что было **в сарайчике Безумного Билли...**







...было бы для нас **невероятно полезным.**







На самом деле **НЕТ** такого распределения выборки, на которое можно было бы взглянуть!

Это все плод моего воображения.

Я помню каждую консервную банку, которую когда-либо продавал.

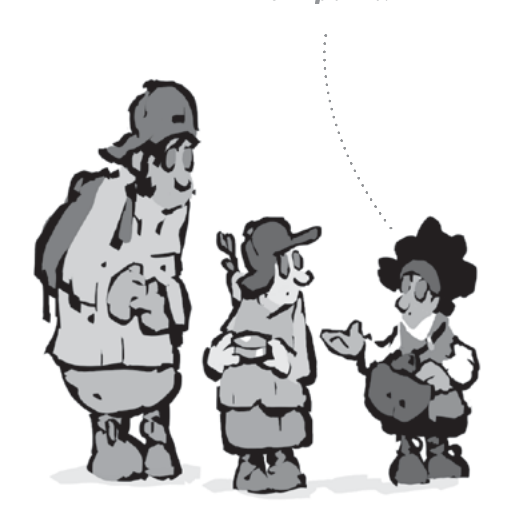
Как показывает практика...

...все, что мы можем получить, это **одну банку.**




Ну, так что, какова средняя длина всех червей в болоте?
Эм-м...

Глава 9
СТАТИСТИЧЕСКИЙ ВЫВОД





Ясно как день, что у нас по-прежнему есть нерешенная проблема...



*...и она сводится к **следующему:***

мы пытаемся обнаружить нечто, ***чего не можем увидеть.***

Невозможно, заглянув в одну выборку...

...увидеть среднее значение в генеральной совокупности.

Мы всего лишь 30 жалких червяков, в то время как в болоте можно найти еще сотни тысяч миллионов.

Как будто мы бредем в тумане на ощупь, **пытаясь найти снежного человека.**

Я **верю** всем сердцем, что он где-то там!

Но все равно вы **никогда** не сможете его найти.

К счастью, хотя мы и **не можем увидеть** то, что ищем...

Ничего не вижу **за туманом.**

...мы **можем** продолжать искать **подсказки...**

...которые помогут нам понять, где то, что мы ищем, **может находиться.**

Если бы ты был средним значением совокупности, **где бы ты прятался?**

Под куполом того холма!

Когда мы пытаемся угадать местонахождение среднего значения генеральной совокупности...



...мы можем опираться в своем предположении на что-то, в чем уже уверены...



...и мы с вами даже уже выяснили, что это такое:







Вот что мы сейчас будем делать:

поскольку **средние значения выборки** обычно **скапливаются...**

...вокруг среднего значения генеральной совокупности...

...вроде этого...

...мы можем **нарисовать вот такую возвышенность...**

...чтобы **угадать, где находится среднее значение генеральной совокупности.**



Полагаю, оно **под куполом этого холма!**

Статистики называют этот процесс статистическим выводом...

Мы не можем увидеть его собственными глазами...

...поэтому ищем те значения, которые, как нам кажется, будут **концентрироваться** вокруг него.

Это похоже **на охоту за снежным человеком...**

...когда выходишь на след, обнаружив **отпечатки огромного размера.**



...мы собираемся посвятить всю оставшуюся главу тому, чтобы в общих чертах обрисовать первый шаг.

Мы хотим использовать **одну выборку...**



...чтобы **представить себе, что бы мы увидели, если бы отобрали гораздо больше экземпляров для исследования.**





...создание иллюстрации.



На этой картинке видны, как нам кажется, **средние значения выборки...**

...если бы мы пошли и собрали сто тысяч миллионов образцов.

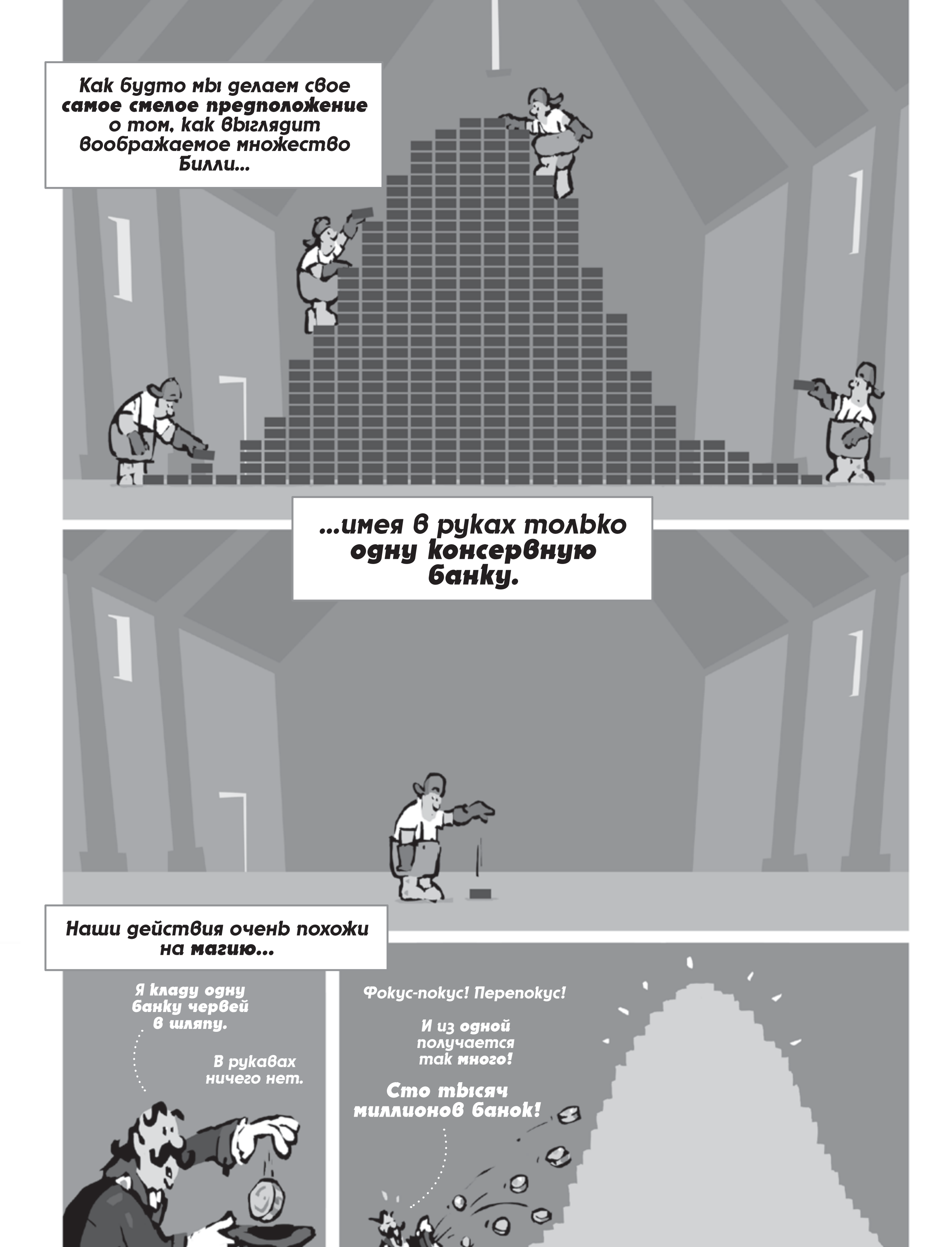







И в основе всего этого лежит информация, которую мы получаем из одной **случайной выборки.**




Как будто мы делаем свое самое смелое предположение о том, как выглядит воображаемое множество Билли...
...имея в руках только одну консервную банку.
Наши действия очень похожи на магию...
Я кладу одну банку червей в шляпу.
В рукавах ничего нет.
Фокус-покус! Перепокус!
И из одной получается так много!
Сто тысяч миллионов банок!
...Хотя на самом деле все очень просто.

Чтобы сделать нужный рисунок, мы используем центральную предельную теорему в качестве плана:



1. Помните, что тут возникают некоторые ограничения, см. стр. 102.

Поскольку мы не знаем настоящих значений в генеральной совокупности...



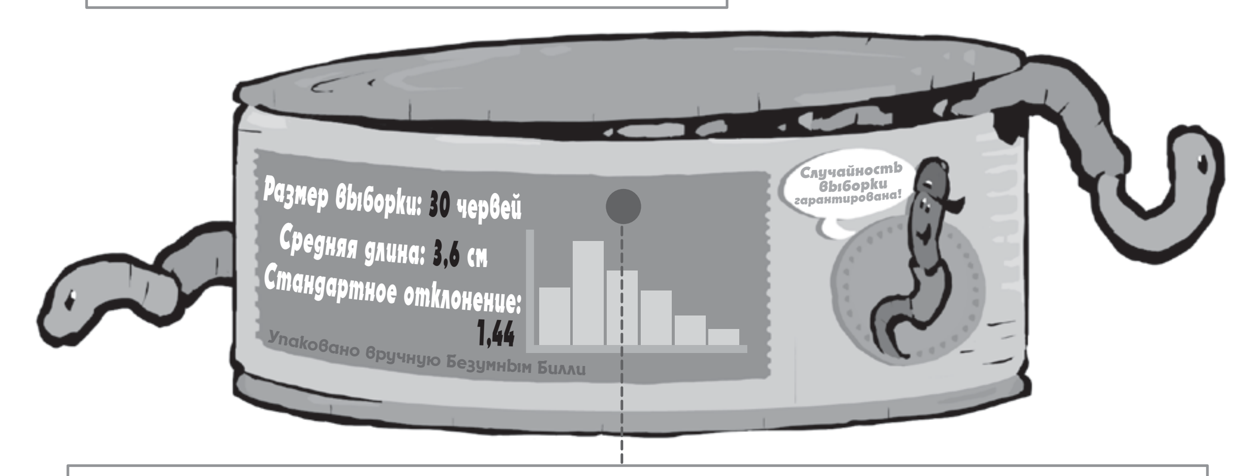
...мы просто заменяем их теми, которые получили из нашей выборки.







Так, например, когда мы используем выборочные значения из одной банки...



...мы рисуем картинку, которая выглядит примерно так:

3,6

Наше **предполагаемое** огромное множество средних значений **нормально распределено...**

...и **центрировано по среднему значению в одной консервной банке...**

...со **стандартным отклонением,** равным...

...**нашему стандартному отклонению,** поделенному на квадратный корень объема выборки!

Вот так!

$$CO = \frac{1{,}44}{\sqrt{30}}$$

Мы называем эту картинку предполагаемым выборочным распределением*.





Это предположение...

...о том, как средние значения выборки могли бы быть **распределены...**

...если бы мы насобирали их целую тонну.





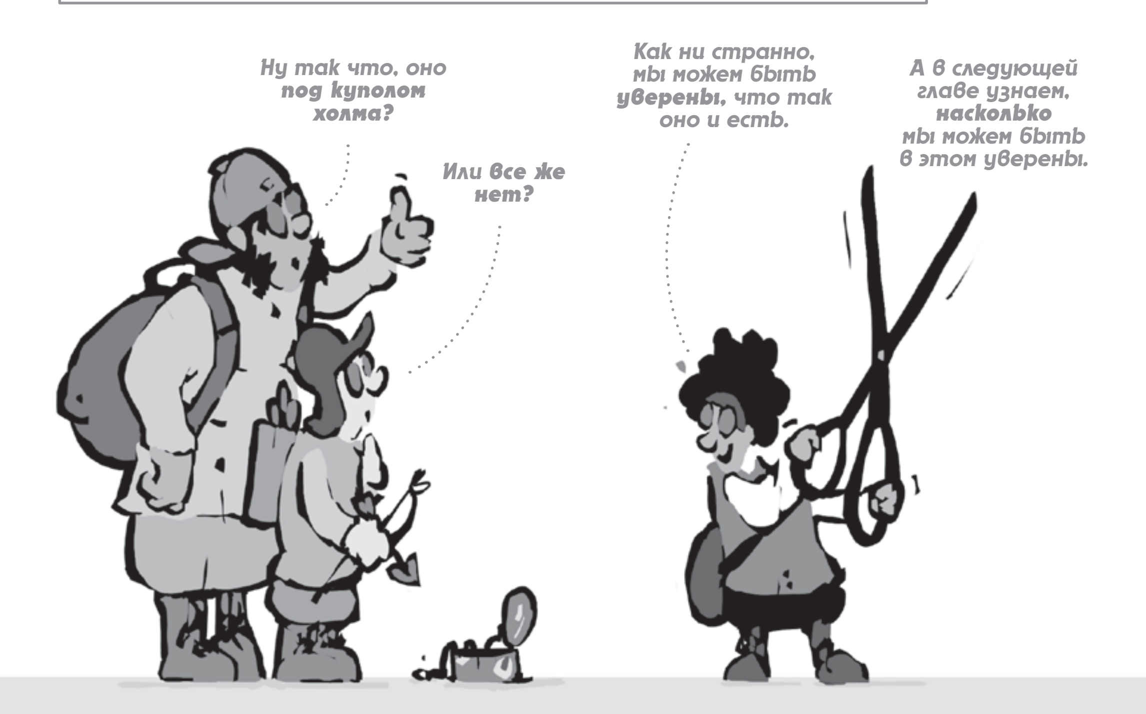
* См. стр. 219, чтобы узнать, как описывать подобные случаи, используя математические символы.

Теперь, использовав **одну выборку...**

...чтобы создать **предварительное распределение выборки...**



...мы можем **подвести итоги нашей охоты на среднее значение генеральной совокупности.**

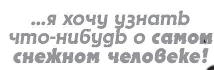


Глава 10
ДОСТОВЕРНОСТЬ



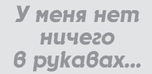


Помните, что в результате нам нужно узнать что-нибудь о среднем значении в генеральной совокупности.





К сожалению, несмотря на все те **магические трюки,** которым мы только что научились...



...мы **никогда** не сможем **добиться этого.**



Никогда!



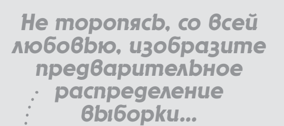
Вот поэтому мы и учимся делать **предположения.**





Мы пока поговорили только о том, как выглядит первый шаг в **процессе выстраивания предположений...**

...но нам еще предстоит разобраться со вторым.







Итак, в этой главе мы научимся **детализировать наш рисунок...**



...состригая аккуратненько по краям...

...и используя то, что осталось, чтобы **вычислить степень достоверности.**

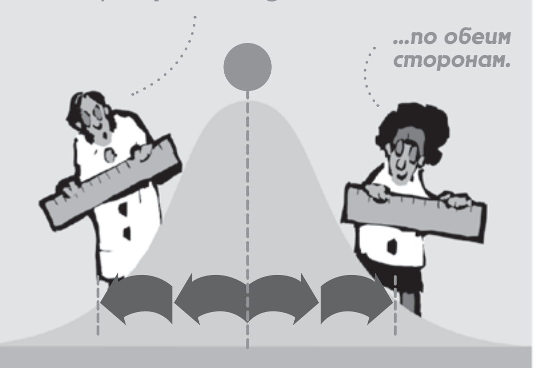


Раз уж мы уже знаем, как нарисовать предварительное распределение выборки...



...мы легко научимся высчитывать и степень достоверности.

Нужно просто хорошенько вглядеться в то, что мы только что нарисовали...

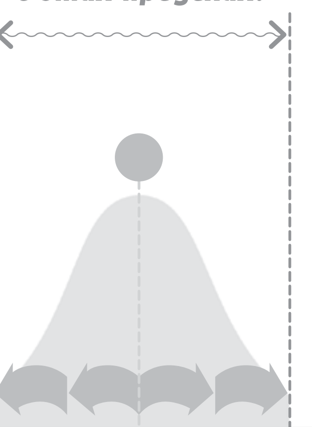


...и отрезать все «хвостики»!



А потом мы делаем утверждение, например такое:







Только и всего!



Если мы хотим **большей достоверности**, нам нужно просто **отрезать чуть дальше**.

Если мы хотим быть **уверены** на 99,7%...

...то мы отрезаем в **трех стандартных отклонениях** от центра...

...с каждой стороны.

А если нам нужно **меньше достоверности**, то нужно **отрезать чуть ближе**.

Если мы хотим быть **уверены** на 68%...

...то мы отмеряем и отрезаем в **одном стандартном отклонении** от центра...

...с обеих сторон.



В основе всех этих подсчетов лежит то, что мы уже изучили на стр. **115**!

Но где бы мы ни отрезали, мы **всегда** декларируем нашу степень уверенности с помощью двухчастного утверждения...

Мы уверены на 95%...

...что среднее значение в совокупности находится где-то в этом пределе!

...в котором объединены и **степень достоверности**...

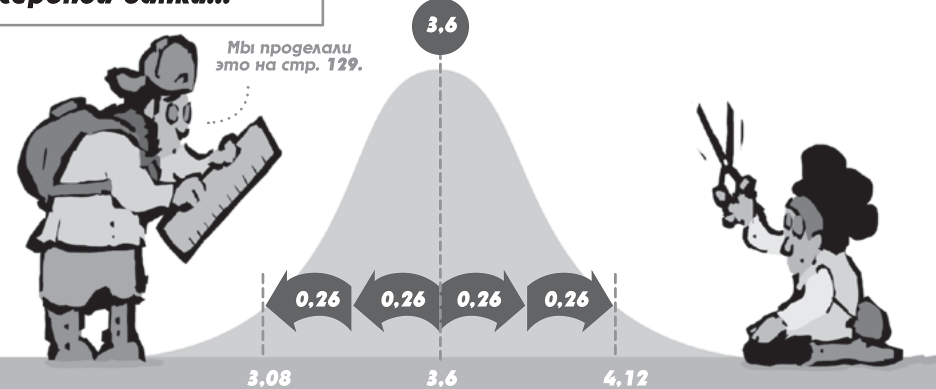
...и **доверительный интервал***.





* Доверительный интервал представляет собой **интервальную оценку**. См. стр. **220**, если хотите узнать больше.

Например, если мы возьмем предварительное выборочное распределение, сделанное с помощью нашей консервной банки...



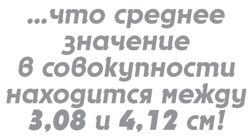
...и отрежем «хвостики»...





...мы сможем сказать следующее:

Мы уверены на 95%...







Но что конкретно это означает?

Мы создали всего лишь **одно** предварительное выборочное распределение с помощью одной случайной выборки...

Мы построили это с помощью одной банки...

3,6

...наполненной 30 червями.

0,26

...и использовали его, чтобы подсчитать **один** доверительный интервал.



Но если бы мы взяли **другую случайную выборку** и использовали ее, чтобы создать другое предположительное распределение выборки...



...мы бы, скорее всего, получили **другой интервал!**



И **если** бы мы продолжали собирать **новые и новые** случайные выборки и выстраивать **новые и новые** предполагаемые выборочные распределения...

Собери и подсчитай.

Собери и подсчитай.

Собери и подсчитай.

Собери и подсчитай.

Собери и подсчитай.

...мы бы продолжали получать **разные интервалы.**

Собери и подсчитай.

Собери и подсчитай.

Собери и подсчитай.

Собери и подсчитай.

Это важно, потому что **единственный вывод**, который мы можем сделать из этого...



Мы на 95% уверены...

...что среднее значение совокупности находится где-то в этом пределе!

...это что, если бы мы установили **таким образом** сто тысяч миллионов разных пределов...



...среднее значение генеральной совокупности **содержалось бы** примерно в 19 банках из 20...

В 19 банках из 20. Получается **95%!**

А примерно в 1 банке из 20 было бы по-другому.

В одной банке из 20 все **совершенно неправильно!**

Будем надеяться, что мы не взяли именно **ту** банку.

Другими словами, когда мы говорим *так:*

мы уверены на 95%...

...что среднее значение генеральной совокупности находится где-то в этом пределе!

...это означает, что есть 5%-ная вероятность, что мы заблуждаемся на сей счет.

В таком случае среднее значение генеральной совокупности **на самом деле где-то в другом месте...**



Печальная правда заключается в том, что **любая выборка**, которую мы отобрали случайным образом из генеральной совокупности...





...может оказаться обманчивой и ввести в заблуждение.





А если **одна наша выборка** дает настолько неоднозначные результаты...

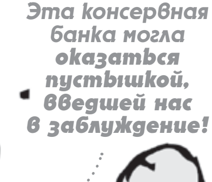
...то и предварительное выборочное распределение, которое мы рисуем на ее основе, **тоже окажется неверным.**





Это **серьезная проблема...**

...но мы можем избежать подобных трудностей, если **всегда** будем помнить о **более масштабной картинке.**







Даже если одна выборка оказалась обманчивой...





...в долгосрочной перспективе станет ясно, что, скорее всего, это не так...

...потому что **большинство средних значений в случайных выборках** имеют тенденцию группироваться вокруг среднего значения генеральной совокупности!



Иными словами, средний показатель одной банки **может** случайно оказаться здесь...



...или **здесь**...



...а такое **маловероятно**...



...и мы можем **быть в этом уверены.**



Подводя итог, скажем, что понимание статистической достоверности...







...предполагает, что мы должны держать в уме как продолжительный период, так и короткий промежуток одновременно.



В долгосрочной перспективе наша приблизительная оценка и отсечение «хвостиков» дают прекрасные результаты. И точка.





...Но вот в краткосрочной перспективе всегда есть вероятность, что мы схватили не ту банку!



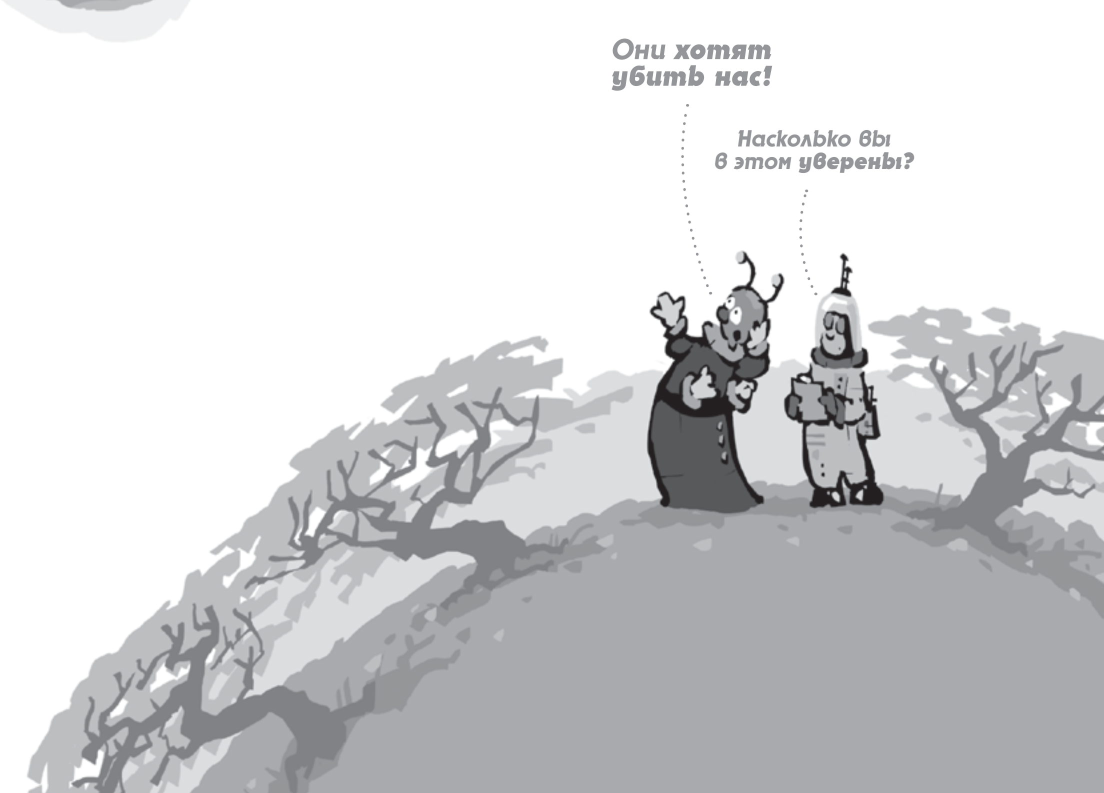




Глава 11

ОНИ НАС НЕНАВИДЯТ





Когда мы используем только одну **выборку...**

Если взять за основу 50 случайно отобранных русалок...

...чтобы **подсчитать степень статистической достоверности** всей генеральной **совокупности...**

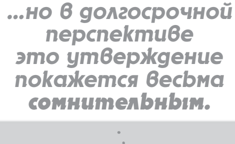


...мы делаем что-то **удивительное!**

Мы делаем **предположение...**

...заслуживающее доверия предположение...

Кто бы мог подумать, что русалки **такие крошечные?**



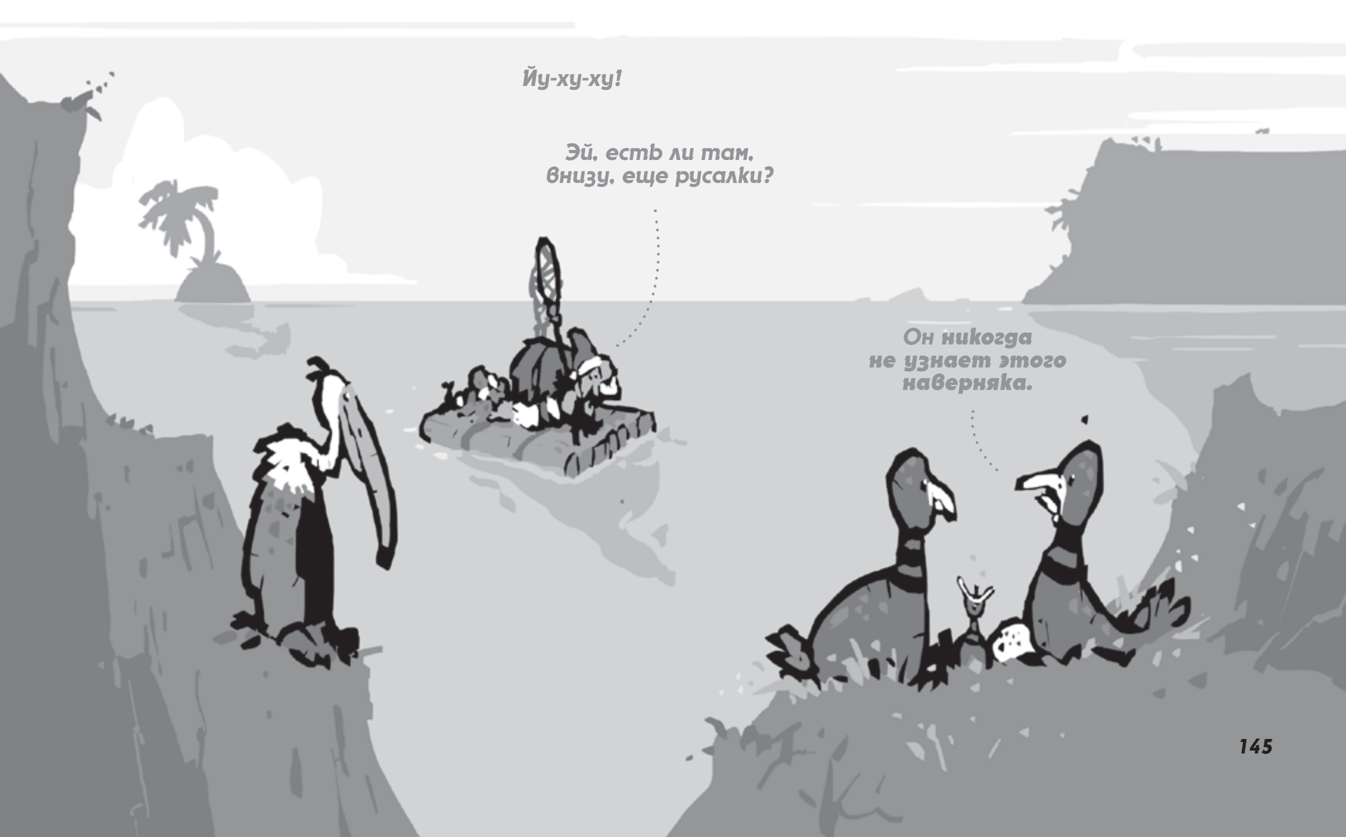
Одна случайная выборка может сильно запутать...

...но в долгосрочной перспективе это утверждение покажется весьма **сомнительным.**





...о чем-то, что мы не можем увидеть, а способны только представить себе.




Все начинается
с трех цифр.
Достаточно
большой объем
выборки...
...среднее значение
выборки...
...и стандартное
отклонение
выборки.
Но помните,
что все получится только
в том случае, если все
ваши русалки будут
отобраны произвольным
образом.



...и отрезать «хвостики»...





7

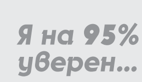
10

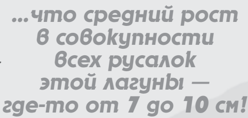
...чтобы получить единственное заслуживающее доверия утверждение...

В долгосрочной перспективе это отлично работает!

...в котором будут и степень уверенности...

...и доверительный интервал.







Как мы уже знаем, из-за того что этот метод требует изрядного количества* подсчетов*...

Ну во-о-от!



...он хорош только для* тех характеристик*, которые* можно измерить*.



Поэтому* может показаться*, что он неприменим к характеристикам, которые не поддаются* явному численному измерению*...

Они счастливы?

Все ли они поют красиво?

Если я ткну в них палочкой, насколько им будет больно?

***...однако* это не всегда так.**

Правда заключается в том, что мы можем высчитать степень достоверности **относительно любой характеристики...**



...если найдем способ **измерить** ее...



...и сможем отметить показатели на **числовой оси.**



В этой главе мы будем заниматься как раз этим...

...чтобы исследовать вопрос, касающийся **ненависти.**



Всем известно, что негодники, живущие на планете **Бип**...

...**ненавидят** хороших людей, живущих на соседней планете **Пип**.

Фу-у! Фу-у!

Вы, @*$& пипиане!

Они **нас ненавидят!**

Да эта ненависть измеряется **триллионами поколений!**

И вот вопрос, который нас мучает:

...**правда ли это?**

Так как мы не можем опросить лично **все 785 000 000 000 бипиан**, живущих на планете, о том, что они чувствуют...

...нам **ничего не остается**, кроме как основываться в своих суждениях на случайной выборке.

Бипиан слишком много.

Иными словами, нам **придется воспользоваться статистическими приемчиками.**

Помните, никто не использует статистические техники, **если только в этом нет нужды!**

Но прежде чем мы отправимся делать случайную выборку...



В этом случае мы можем придумать свою систему исчисления...



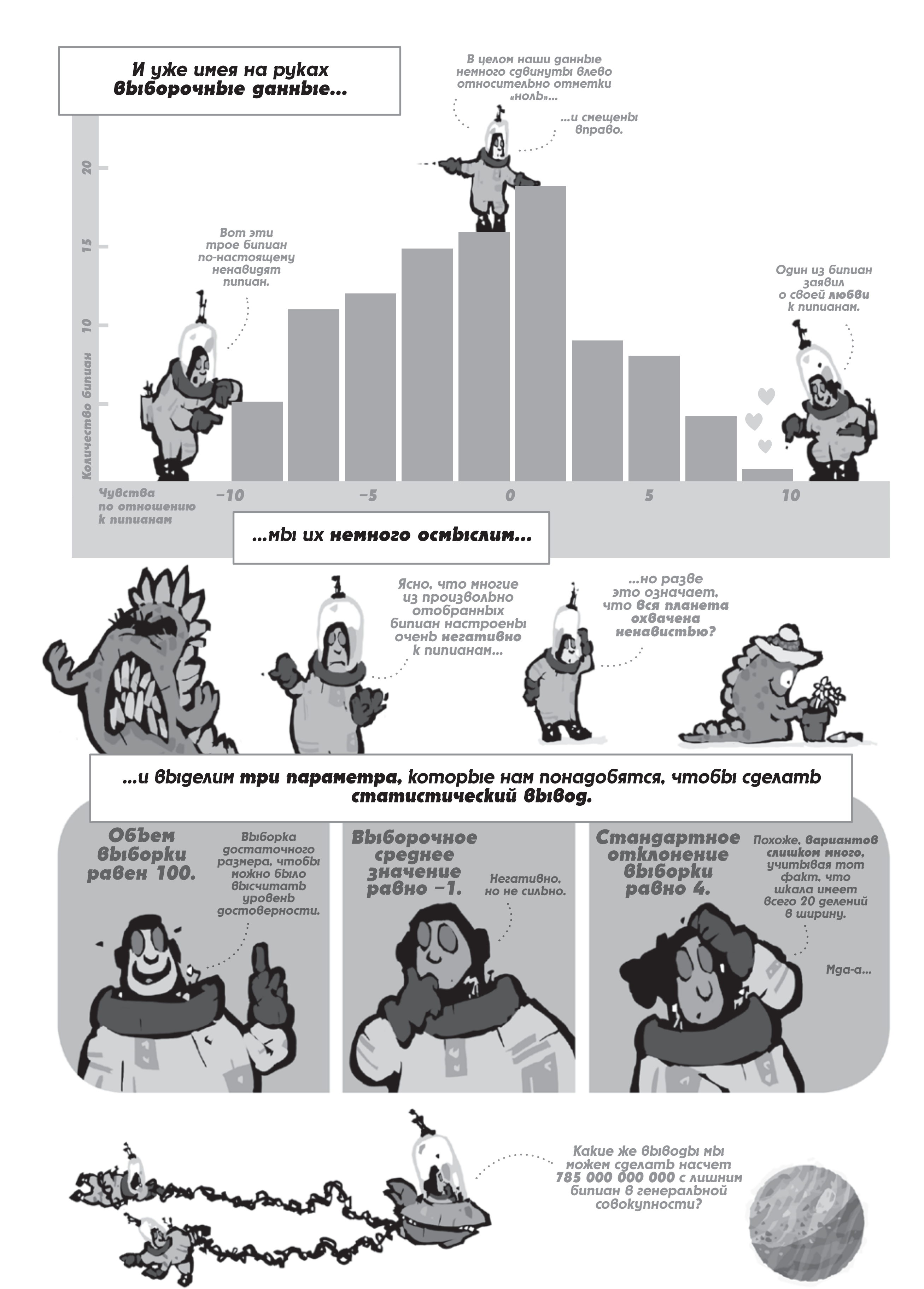
Давай переведем слова каждого бипианина, с которым поговорим...

...в числовой эквивалент по шкале от −10 до 10.

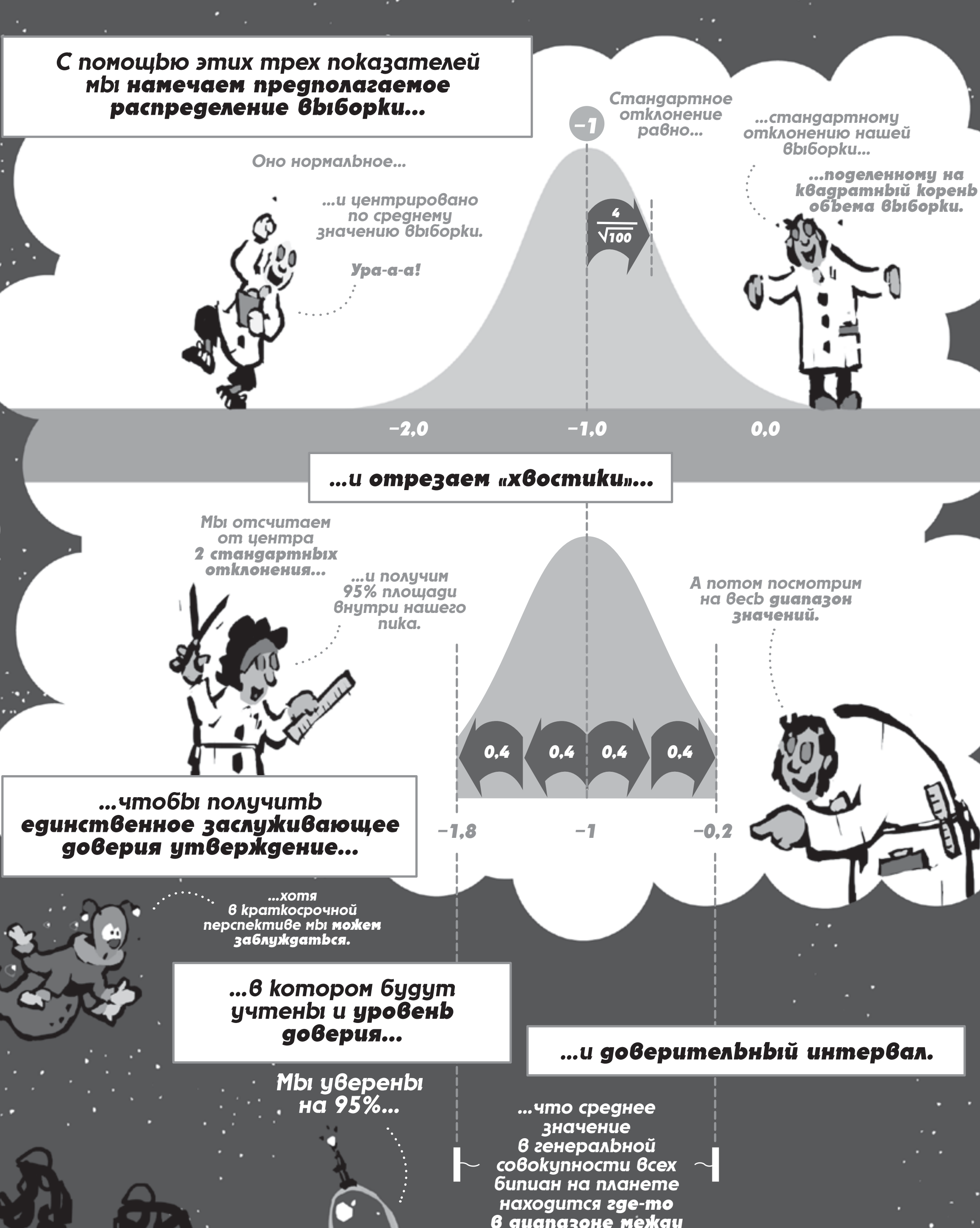



Как бы вы оценили свою ненависть к пипианам по шкале от −10 до 10?
Как бы вы оценили свою ненависть к пипианам по шкале от −10 до 10?
Как бы вы оценили свою ненависть к пипианам по шкале от −10 до 10?
Затем мы аккуратно соберем все данные, которые...
...получили произвольным образом...
Вообще, между выборкой бипиан, которых мы опрашиваем...
...или любой другой выборкой...
...может и не быть никакой систематической разницы...
...поэтому давайте-ка поищем по всей планете и отберем бипиан совершенно случайным образом...
...немного отсюда...
...немного отсюда...
...нескольких отсюда...
...чуть-чуть отсюда...
...немного отсюда...
...и здесь поищем...
...и т. д.
...и т. д.
...памятуя о том, что собрать нам нужно достаточное количество...
...чтобы быть уверенными, что все подсчеты верны.


...и не остановимся, пока не отберем 100 бипиан.


И уже имея на руках выборочные данные...
В целом наши данные немного сдвинуты влево относительно отметки «ноль»...
...и смещены вправо.
Вот эти трое бипиан по-настоящему ненавидят пипиан.
Один из бипиан заявил о своей любви к пипианам.
Количество бипиан
20
15
10
Чувства по отношению к пипианам
−10
−5
0
5
10
...мы их немного осмыслим...
Ясно, что многие из произвольно отобранных бипиан настроены очень негативно к пипианам...
...но разве это означает, что вся планета охвачена ненавистью?
...и выделим три параметра, которые нам понадобятся, чтобы сделать статистический вывод.
Объем выборки равен 100.
Выборка достаточного размера, чтобы можно было высчитать уровень достоверности.
Выборочное среднее значение равно −1.
Негативно, но не сильно.
Стандартное отклонение выборки равно 4.
Похоже, вариантов слишком много, учитывая тот факт, что шкала имеет всего 20 делений в ширину.
Мда-а...
Какие же выводы мы можем сделать насчет 785 000 000 с лишним бипиан в генеральной совокупности?


Итак, еще раз, для начала нас интересуют три показателя.
Объем выборки, которую мы сочли достаточно большой, равен 100...
...выборочное среднее значение равно −1...
...а стандартное отклонение выборки равно 4.
Если мы не можем сделать выборку достаточного объема...
...нам придется воспользоваться специальными инструментами, о которых мы узнаем в главе 14.



С помощью этих трех показателей мы намечаем предполагаемое распределение выборки...
Оно нормальное...
...и центрировано по среднему значению выборки.
Ура-а-а!
−1
Стандартное отклонение равно...
$\frac{4}{\sqrt{100}}$
...стандартному отклонению нашей выборки...
...поделенному на квадратный корень объема выборки.
−2,0
−1,0
0,0
...и отрезаем «хвостики»...
Мы отсчитаем от центра 2 стандартных отклонения...
...и получим 95% площади внутри нашего пика.
А потом посмотрим на весь диапазон значений.
0,4
0,4
0,4
0,4
...чтобы получить единственное заслуживающее доверия утверждение...
−1,8
−1
−0,2
...хотя в краткосрочной перспективе мы можем заблуждаться.
...в котором будут учтены и уровень доверия...
...и доверительный интервал.
Мы уверены на 95%...
...что среднее значение в генеральной совокупности всех бипиан на планете находится где-то в диапазоне между −1,8 и −0,2!

Итак, что же?

Ненавидят ли бипиане пипиан?

Руки чешутся выдавить им глаза пальцами...

...отрезать бы им их хвостики тупым ножом...

...ну-ка, выплесните всю свою злобу...

...и подожгите пары своей ненависти!

Видел?

Просто какая-то кучка **озлобленных извергов!**

Раз уж мы не можем опросить всю генеральную совокупность, мы **никогда не получим точный ответ на свой вопрос.**

Но вот наши выборочные данные...

...весьма однозначно говорят о том, что не ненавидят!

Мы уверены на 95%...

...что настоящее среднее значение в генеральной совокупности находится не где-то там, в зоне ненависти...



–1,8

–0,2

А тут мы, в лучшем случае, наблюдаем мягкое неприятие...

...граничащее с безразличием...

...но **не ненависть!**

–10

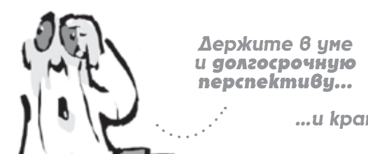
–5

0

5

10

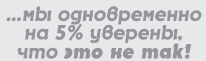
Помните о том, что максимум, который мы можем предложить с помощью статистики, это портрет с некоторыми нюансами...



...потому что **любое заключение, сделанное на основании данных выборки...**



...**может быть в корне неверным.**

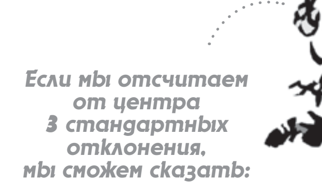


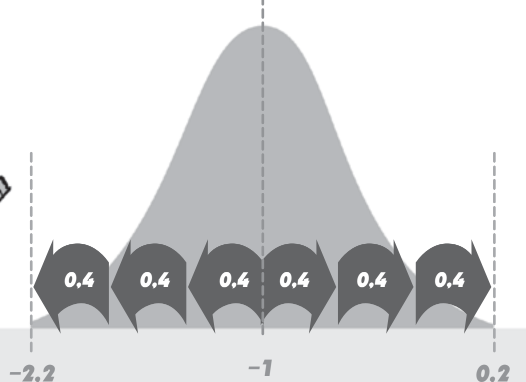




И даже если мы **расширим наш уровень** достоверности...

...чтобы охватить **больший** интервал...











...все равно **есть вероятность,** что мы ошибаемся.





Мы только что высчитали доверительные интервалы, равные **95%** и **97,7%**...

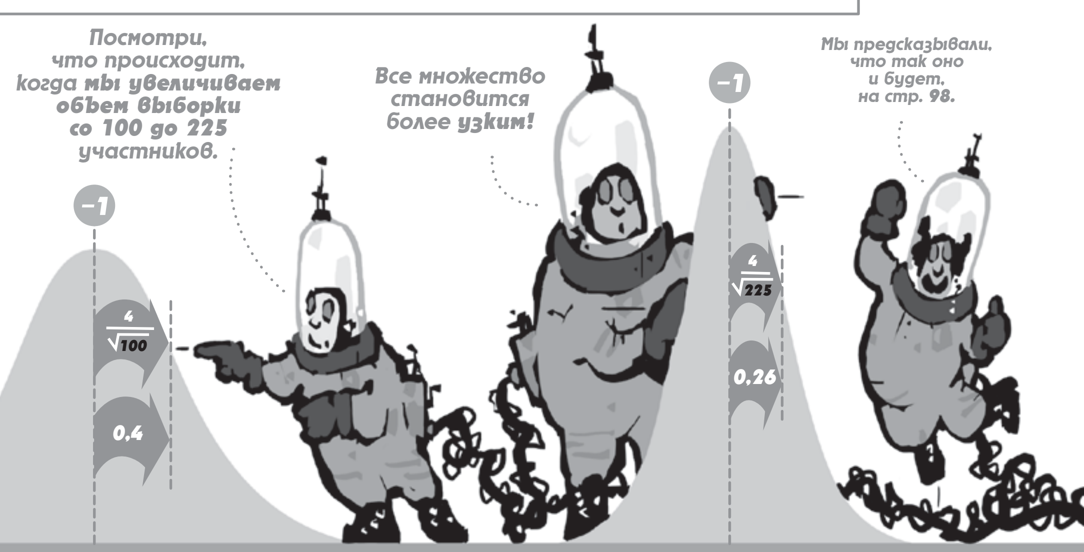
...основываясь всего на одной произвольной выборке **в 100 бипиан.**



Если бы мы изначально опросили **больше случайно отобранных бипиан...**



...наше предварительное выборочное распределение могло оказаться **более узким...**



...а это, в свою очередь, **сузило бы** наши доверительные интервалы, а значит, и сделало их **более точными!**



Чтобы посмотреть, как это работает, вспомните, что мы начали с **этих трех показателей.**



Объем выборки равен **225.**

Бипиан стало больше в два с лишним раза.



Выборочное среднее значение равно −1.

Стандартное выборочное отклонение равно 4.

Маловероятно, что объем нашей выборки, а также стандартное отклонение останутся такими же и в большей выборке...

...но давайте возьмем то же число и посмотрим, что будет, если **мы поменяем только объем выборки.**



Так же, как и раньше, с помощью этих трех цифр мы можем наметить **предполагаемое распределение выборки...**

...и **отрезать «хвостики»...**



Это нормально...

...и центрировано по нашему выборочному среднему значению...

...но в целом на этот раз все **у́же,** потому что объем нашей выборки был больше.

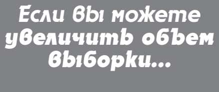
Давайте отсчитаем 3 СО от центра и получим доверительный интервал в 99,7%.

И на этот раз, **какой бы конкретно доверительный уровень** мы ни выбрали...

...у него будет **намного более точный интервал.**



На этот раз мы уверены на **99,7%...**



...что среднее значение в совокупности находится **где-то между −1,78 и −0,22.**

Это почти то же самое, что и **95%-ный** доверительный интервал, который мы получили при опросе **100** бипиан!

Вот, собственно, почему **больший объем выборки всегда лучше!**



Если вы можете **увеличить объем выборки...**

...сделайте это.

Это только усилит вашу уверенность!



В этой главе мы создали **числовую шкалу**, благодаря которой можно определить, **как одна группа людей относится к другой.**

Мой ответ был оценен в 10 баллов по предложенной шкале!

Ты выйдешь за меня замуж?

Эту хитрость можно использовать, чтобы получить ответы **на самые разнообразные вопросы...**







...потому что, если мы соберем достаточно **числовых выборочных данных...**

...мы сможем **высчитать степень достоверности** в отношении **любой совокупности...**

Ты считаешь меня симпатичным?





...находящейся вне пределов досягаемости.

Глава 12

ПРОВЕРКА ГИПОТЕЗ





...намечая предполагаемое распределение выборки...



...и вырезая большой массив в самом центре этого распределения.

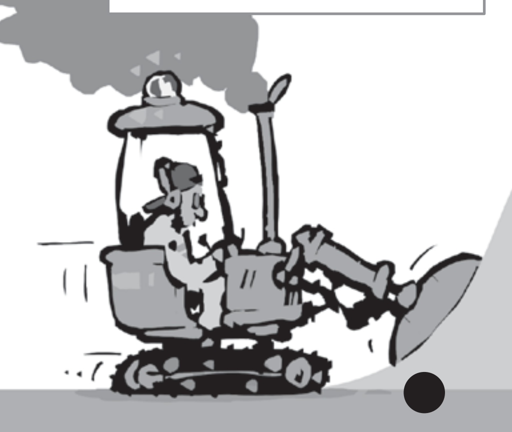






***В этой главе мы познакомимся с* новой техникой.**

Мы возьмем ***нашу предварительную оценку...***



...и посмотрим, что она может сказать, если мы ***сдвинем ее на новое центральное место.***



*Все это часть процесса, который носит название «**проверка гипотезы**»...**

Это такой ***тест...***

...который показывает, правда ли мы думаем, что среднее значение генеральной совокупности ***может быть...***

...ПРЯМО ЗДЕСЬ!





...и является еще одной ***важной стратегией для формирования статистических выводов.***

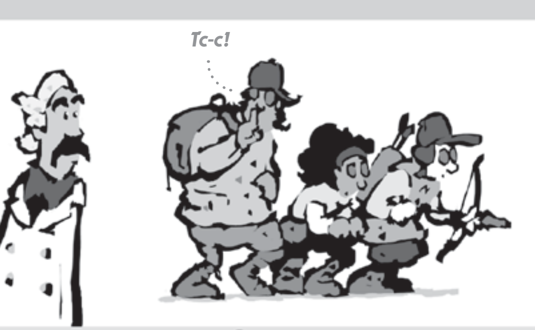
* *См. стр. 221, если вас интересует более подробное описание того, о чем мы говорим в этой главе.*

«Проверка гипотезы» звучит забавно...





...но на деле это просто еще один способ **поохотиться на неуловимое среднее значение генеральной совокупности.**



Как мы уже знаем, нам **никогда** не удастся увидеть среднее значение в генеральной совокупности своими глазами...



...но оказывается, мы можем немного продвинуться в нашем квесте, **попытавшись определить его точное местонахождение.**







Что мы в итоге будем делать с проверкой гипотез?

Проверять наши догадки...

...сравнивая их со средним значением в выборке, **которое мы на самом деле нашли.**







Вообще, вся эта проверка подразумевает, что нужно будет еще не раз **покрутить** и повертеть найденные значения...





...поэтому **спешить не будем.**



Процесс начинается с того, что мы намечаем одно предполагаемое распределение с помощью одной выборки...

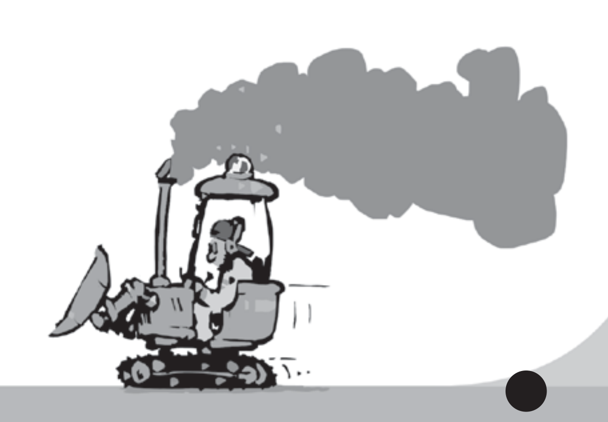


...и передвигаем его в другое место, про которое как раз и хотим все узнать.



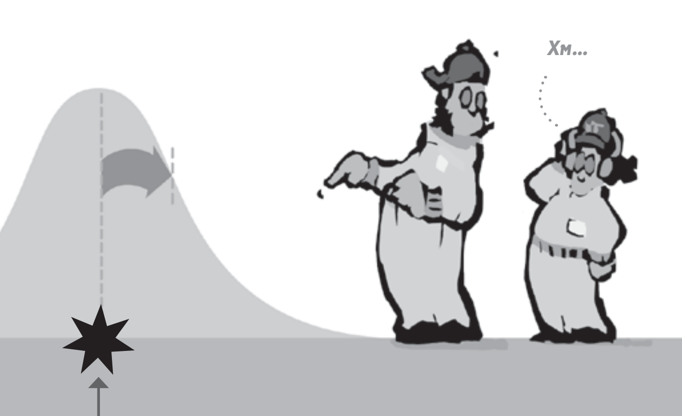
Благодаря этому мы можем увидеть, как выглядели бы другие выборочные средние значения...

...если бы они были собраны воедино и центрированы в этом месте.





Затем мы **смотрим** на среднее значение, которое нашли в нашей выборке, и решаем:



если бы среднее значение генеральной совокупности действительно находилось где-то здесь...

...какова вероятность того, что мы бы в произвольном порядке нашли такую выборку, как наша?

Чтобы узнать ответ, мы можем воспользоваться теми знаниями, которые получили о **статистической достоверности!**



Если наше предположение верно и настоящее среднее значение генеральной совокупности **находится здесь...**

...тогда в **долгосрочной перспективе** мы можем предполагать, что в любой выборке...

...среднее значение будет находиться **под куполом холма.**



Но если среднее значение в нашей выборке будет отличаться **в одну сторону...**



...или **в другую...**



...мы сможем резонно предположить, что наши гипотезы **оказались ложными.**





Но давайте все же **поконкретнее** об этом.

В долгосрочной перспективе мы предполагаем, что 95% всех средних значений в выборке будут находиться **в пределах двух стандартных отклонений от среднего значения генеральной совокупности...**

...поэтому вероятность того, что мы наугад возьмем среднее значение в выборке **где-то тут...**

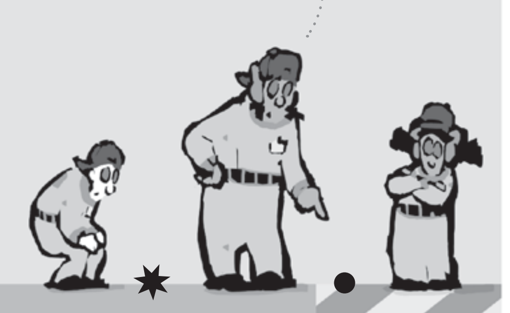
...или **тут...**

...составляет **примерно 5%.**

Ты **можешь** наугад обнаружить среднее значение в выборке где-нибудь здесь...

...но это **очень маловероятно.**

На практике мы сравниваем наши выборку и предположение, подсчитывая то, что называется **вероятностью*** (или Р-значением)...



...и если оно меньше 5%, мы подозреваем, что наше предположение, **возможно, неверно.**




* См. стр. 221-222, где вы найдете словарное определение.

Мы всегда заканчиваем проверку гипотез, ***принимая формальное решение.***



Если выборка и приблизительная оценка оказались ***довольно близко друг от друга...***

Мы получили значение вероятности, равное ***5% или больше,*** *когда сравнили их...*



...мы сделаем вывод, что наша приблизительная оценка, вероятно, была сделана правильно.

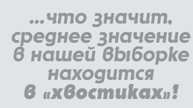
Есть большая вероятность, *что мы бы произвольным образом сделали выборку, похожую на нашу...*

...из генеральной совокупности, ***центрированной прямо там.***





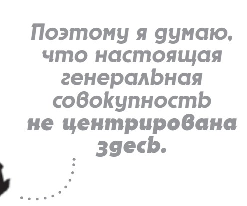
Как бы то ни было, если выборка и предварительная оценка **далеки друг от друга...**













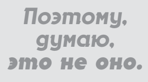


Потому что неважно, каково будет наше формальное заключение...









...есть вероятность, что мы заблуждаемся*!



* Это может показаться довольно нудным, но все это невероятно важные вещи. Подробности на стр. 222.

Помните, проверка гипотез **основывается на оценке**, построенной с помощью одной выборки...



...и если вдруг окажется, что наша выборка **неправильна**...

...выводы тоже будут **неверны**.



К счастью, даже несмотря на то, что мы никогда не можем добиться **абсолютной точности**...





...проверка гипотез может порой оказаться **очень полезной**.

Как мы увидим в следующей главе, все эти **тонкости**, **возникающие при проверке гипотезы...**



...становятся особенно актуальны, когда мы пытаемся ответить на **один конкретный** вопрос:

будет ли доказательство **достаточно сильным...**

...чтобы заставить нас **усомниться** в этой гипотезе?

Глава 13
ПРОТИВОСТОЯНИЕ



Теперь, изучив с формальной точки зрения, какие шаги предпринимают для проверки гипотезы...

...давайте посмотрим, как это происходит в жизни.





На практике мы используем все эти шаги...

...чтобы столкнуть одну идею......



...с другой.



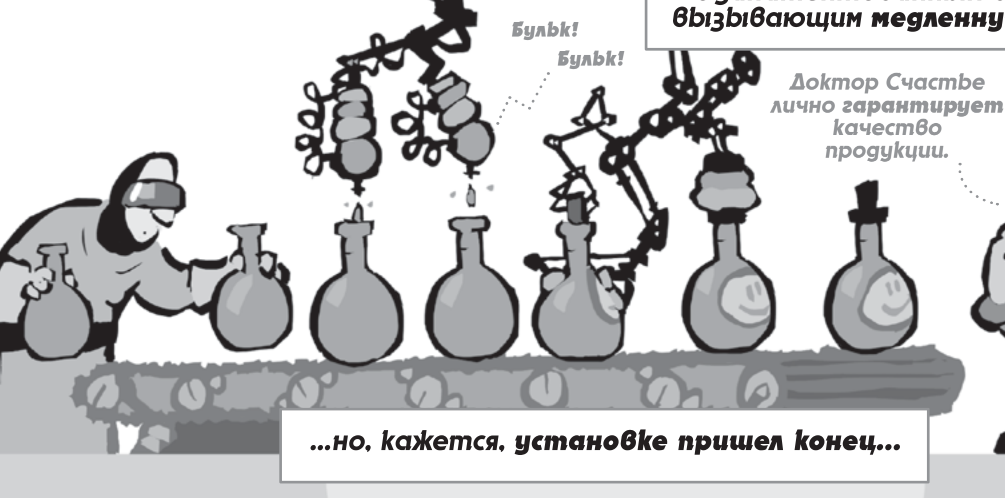



Все это похоже
на **реслинг...**


...но чтобы победить, здесь
все нужно делать **наоборот.**
Побеждает
всегда самая
неоригинальная
идея...


...если только наше
доказательство
не окажется
достаточно
сильным, чтобы
перевесить!
Чтобы своими глазами увидеть, как это работает,
давайте разберем несколько **историй.**


Представьте себе, что **Минерва Хайтауэр** (**назовем ее Доктор Счастье**)...
Привет!
Я самый главный в мире специалист по **козням!**
...переживает из-за своей **установки, в которой готовится яд.**


Она должна разливать в среднем по **0,25 г чистого зла...**
...в каждую из этих склянок с запатентованным бальзамом, вызывающим **медленную смерть...**
Бульк!
Бульк!
Доктор Счастье лично **гарантирует** качество продукции.
...но, кажется, **установке пришел конец...**


Хозяйка, **клиенты жалуются...**
...некоторые уверяют, что в бальзаме **недостаточно зла...**
Но, дорогой мой, у меня **репутация!** И я должна за ней **следить!**




...а другие говорят, наоборот, его **слишком много.**
Какое **расточительство!**


...поэтому Минерва начинает **проверку.**

Итак, Доктор Счастье делает **случайную выборку...**





...сортирует...

...и, сделав подсчеты, узнает среднее значение: **0,14 г.**



Оказывается, в каждой склянке **недостаточно** чистейшего зла.

Должно быть, сломалась моя старушка!

Но вот ведь **проблема:**

Доктор Счастье **вознамерилась купить новую установку!**



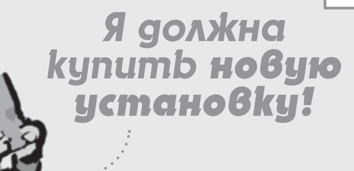
Но ее бухгалтер не дает это сделать, пока они оба не **убедятся,** что установку уже **действительно не спасти.**



Чтобы **оценить достоверность,** Доктор Счастье может **воспользоваться проверкой гипотез.**




И во время проверки Доктор Счастье сталкивает между собой эти две идеи:
либо эта старая, повидавшая виды установка отжила свое...
...либо нет.
Одна идея интересная...
...другая неоригинальная.




Я должна купить новую установку!
Только давайте не будем спешить и принимать необдуманные решения, Минерва.
Каждую идею можно объяснить по-своему и увидеть, почему мы получили данные, которые получили...

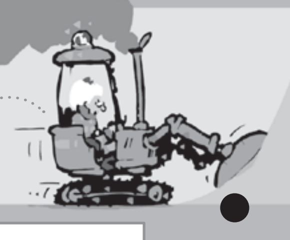


Среднее значение в выборке показывает, что жидкость в склянке какая-то светлая...
...потому что установка просто больше неспособна разливать достаточное количество чистого зла!
Каждая отдельно взятая выборка, вероятно, покажет либо недобор, либо перебор — чисто случайно.
Что, если жидкость в этой выборке бледновата из-за случайной вариации?
...и теперь Доктору Счастье нужно посмотреть на неоригинальное объяснение и подумать, сможет ли она отклонить эту версию!

Поэтому она берет свои данные и намечает предполагаемое распределение выборки.



Затем она сдвигает распределение туда, куда велит неоригинальная идея...



...и высчитывает Р-значение...

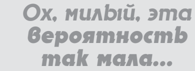


...затем берет время на обдумывание того, что же на самом деле означает полученное Р-значение.



В итоге она со всей уверенностью отвергает неоригинальную идею...

...в угоду интересной.







В этой истории P-значение Доктора Счастье помогло ей принять уверенное решение.



Но помните, в статистике у уверенности всегда есть и обратная сторона.





Поэтому, даже если, кажется, есть доказательства в поддержку решения Доктора Счастье...

...она может ошибаться.





К лучшему это или к худшему, но проверка гипотезы всегда заканчивается именно этим.





Как бы то ни было,
в конце концов Доктор Счастье
получила то, что хотела,
проверив свою гипотезу...



Все, побежала
покупать себе
XT-4300!

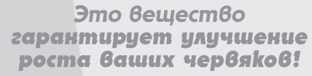
У меня **лучшая**
работа
в мире!

...но так бывает **не всегда.**

А чтобы знать,
как оно бывает,
давайте рассмотрим
еще один случай.



Представьте себе, что Безумный Билли подкармливает червяков в своем болоте стероидами...







...и хочет узнать, эффективны ли они.

И вот он делает случайную выборку...







...сортирует...



...и, сделав подсчеты, получает среднее значение, равное 4,19 см.



Конечно, он знает, что среднее значение в генеральной совокупности червей **раньше было 4 см...**



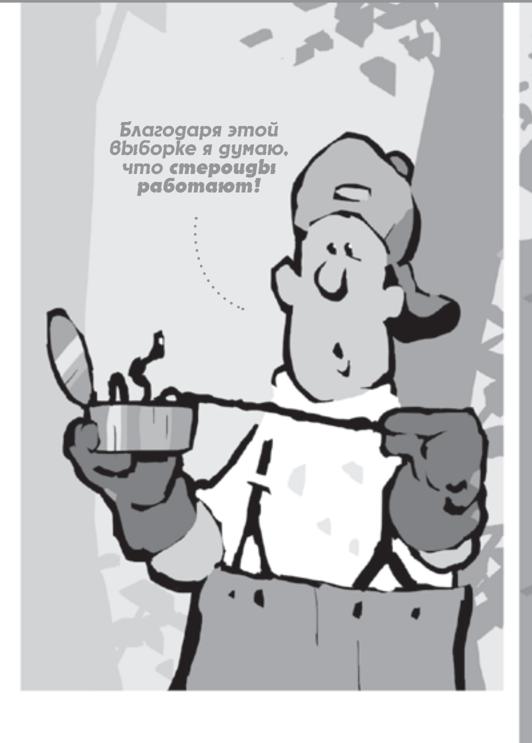
...и **надеется** доказать, что оно и правда **стало больше.**





Но вот ведь **загвоздка:**

у него на руках среднее значение, которое больше **прежнего...**



...но это, возможно, **только благодаря случаю!**



Чтобы решить, прав он или нет, Билли решил **проверить свою гипотезу.**

И он **сталкивает между собой эти две идеи:**





Одна идея **новая и интересная...**

...а другая скучная, **неоригинальная.**





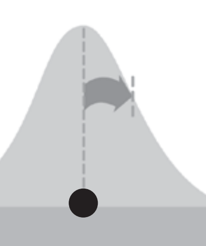
Каждая идея сопровождается **своими объяснениями**, почему мы получили те данные, которые получили...



...и теперь Билли нужно посмотреть на **неоригинальное** объяснение и подумать, **сможет ли он его отклонить!**

И вот Безумный Билли берет свои данные и намечает предполагаемое распределение выборки.





Затем он сдвигает распределение туда, куда велит неоригинальная гипотеза...



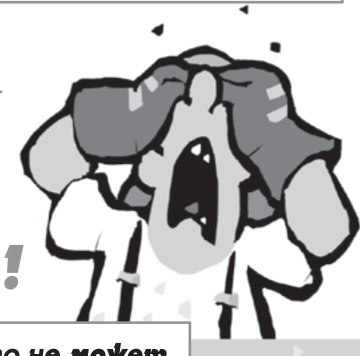
...и высчитывает Р-значение.





Но когда он понимает, что же на самом деле означает его Р-значение...





...он делает неутешительный вывод, что не может быть уверен в эффекте стероидов!

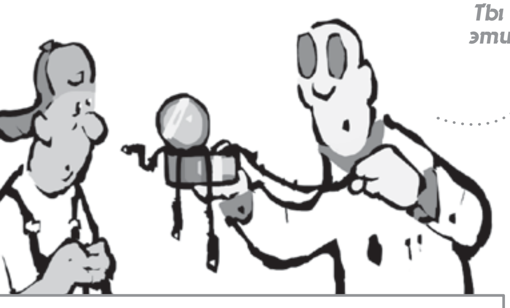





Конечно, заключение, которое сделал Билли...
...не означает, что его стероиды **не работают.**
О черт!
Я не могу быть уверен в том, что мои стероиды работают!


Терпение, **возможно, они все-таки эффективны!**


Ведь в среднем твои червяки оказались длиннее, чем показало предыдущее среднее значение.
Это всего лишь означает, что **это доказательство недостаточно веское,** чтобы подтвердить устраивающее его предположение.


Ты мог получить эти результаты и **наугад!**


Заключение не то чтобы **удовлетворительное...**


Ничего **интересного** тут не происходит.


Мы вернулись к тому, с чего начали.
...тем не менее оно очень **важное.**

Мы рассмотрели два разных примера проверки гипотезы...

Я **жаждала** увидеть доказательство, что моя установка сломалась.

Я **жаждал** доказательств того, что черви стали длиннее.

...и увидели два разных результата.

Я могу быть **уверена** в своем доказательстве. Оно верно.



В обеих историях красивая новая идея...





В обеих историях нам хотелось, чтобы новая идея оказалась правильной...



Весь смысл проверки гипотез в том, чтобы убедиться, что мы не делаем преждевременных выводов.





Глава 14

Летающие свиньи, плюющиеся пришельцы и петарды





Как мы уже видели, обе **стратегии** предполагают **одни и те же основные шаги.**



Теперь, когда мы наконец разобрались с самым **основным**...

Поздравляем!

Если вы понимаете, как **все это** работает...

...вы поняли суть статистики!

Остались **мелочи.**

...мы собираемся посвятить эту главу разбору отдельных **случаев**...

...чтобы дать вам представление о том, что вас ждет, **если вы захотите углубиться в тему.**

Согласно прогнозу, ожидаются **летающие свиньи, плюющиеся пришельцы и петарды!**

Нам, очевидно, понадобится **это!**

Пока что нас в основном интересовало, каковы будут ваши действия в случае, если **УСЛОВИЯ ИДЕАЛЬНЫ:**

мы научились охотиться на среднее значение в **одной генеральной совокупности...**

...используя одну **большую выборку...**

...и **по-настоящему случайные замеры,** полученные при ее изучении.

Водичка чистая...

Солнышко светит...

...Статистика — это **легко!**



На практике же условия бывают гораздо **более неоднозначными...**

А что, если не **удастся собрать выборку достаточно большого объема?**

Что, если вы не сможете сделать замеры, которые будут **действительно случайными?**

А что если вас интересует **что-то другое, а не среднее значение?**

...а **продвинутая статистика** во многом предполагает умение решать сложные задачи.



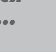


...(а так оно и будет)...

...мы все равно можем полагаться на **основные шаги**, о которых говорили на страницах этой книги...

...нам лишь нужно научиться их **модифицировать.**









Летающие свиньи!

Вот наша первая история: давайте представим себе, что летающие свиньи **в пятнышко** быстрее, чем летающие свиньи **в полосочку**...





...а еще они **намного дороже**...

...и выглядят **омерзительно!**

И поэтому Сэм Нидлхаус хочет узнать, **насколько они быстрее?**

Я запускаю свой **свинобизнес**.

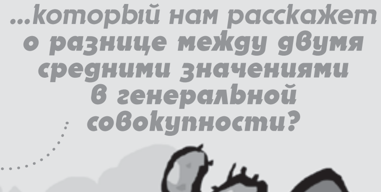
И мне интересно, инвестировать ли сбережения в пятнистых **свинок**...

...или лучше прикупить **полосатых**, а оставшиеся от покупки деньги пустить **на симпатичные костюмчики** для них?



Если говорить языком статистики, вопрос звучит так:







Тогда нам подойдут данные о 40 произвольно отобранных пятнистых свинках...



...и 40 произвольно отобранных полосатых свинках...

Среднее значение в нашей выборке равняется **44,2** км/ч.

Наше СО составляет **4,7** км/ч.

Количество

15

10

30 40 50 60 70

Максимальная скорость (км/ч)

...чтобы наметить предполагаемое распределение выборки...

...которое немного отличается от привычного нам.



Да, теперь распределение выглядит немного иначе, но мы все равно можем его разделить на зоны...*



...и использовать этот метод, чтобы принять уверенное решение.

Куплю, пожалуй, пятнистых!

Если они **настолько** быстрее...

...в долгосрочной перспективе это **окупит** все дополнительные вложения!

* На стр. **223** вы найдете формулу.

ПЛЮЮЩИЕСЯ ПРИШЕЛЬЦЫ!

У гигантских человекоядных жуков-пришельцев с планеты E8M-286 **кислотная слюна...**

Я бы **не стал** это использовать для увлажнения кожи лица!

...и нам интересно, **какова степень ее кислотности?**

Каков **средний PH-фактор их слюны?**

А она **может прожечь бронежилет?**

К сожалению, мы можем поймать и взять пробы только у **нескольких особей...**

...а этого **недостаточно**, чтобы сделать предположения, как нас учили в предыдущих главах.

Мы смогли получить интересующую нас информацию только о **10 случайно отобранных пришельцах...**

...до того, как они **проглотили** нашего последнего специалиста по сбору данных.

Но если объем выборки **меньше 30**, этого недостаточно...

...нам придется прибегнуть к **другим техникам подсчета.**

К счастью, есть одна **замечательная стратегия**, которой мы можем воспользоваться, если проблема лишь в **небольшом объеме выборки.**





Но для начала нам нужно сделать одно **вопиющее в своей дерзости предположение.**

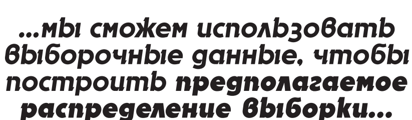
Если мы предположим, что генеральная совокупность нормально распределена...

...(что может быть весьма сомнительно)...



...мы сможем использовать выборочные данные, чтобы построить предполагаемое распределение выборки...

...у которого будет немного более растянутая форма, нежели та, к которой мы привыкли.



Оно, конечно, более растянуто, но мы все равно можем высчитать степень достоверности...*



...просто делать выводы нам нужно особенно осторожно.





* На стр. 223 вы найдете соответствующую формулу.

ПЕТАРДЫ!

Это ужасно, но маленькая Сьюзи Бикер хочет **поиздеваться над соседской кошкой...**



...подложив ей под хвост **петарду!**



Она предпочитает петарды марки **Dingalings...**

...потому что, согласно надписи на упаковке, **у них задержка срабатывания пять секунд...**





Но Сьюзи немного переживает, насколько они **надежны.**







Говоря языком статистики, вопрос Сьюзи сводится к **вариативности...**

...с этим можно разобраться, проанализировав одну выборку...





...чтобы сделать предположение о **размахе вариаций** генеральной совокупности.

В этом случае нам лучше сделать предположение, основываясь на стандартном отклонении вместо среднего значения...

...и весь процесс потребует от нас совершенно других техник подсчета.







Как бы то ни было, мы по-прежнему выполняем самые простые действия:

намечаем немного видоизмененное предполагаемое распределение выборки...*

...и отрезаем «хвостики» вероятности...







...чтобы сделать понятное всем заключение.



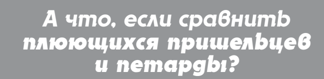




* Все детали на стр. 224.

Как следует из этих историй, нам нужно учитывать огромное количество нюансов...

...когда мы пытаемся ответить на вопросы продвинутой статистики.





И, по правде говоря, нюансам нет числа.

Так, если придется иметь дело с данными, которые как-то друг с другом коррелируют...*

* См. стр. 224, если вас интересуют подробности.





...есть несколько хитростей, которые помогут нам в работе.



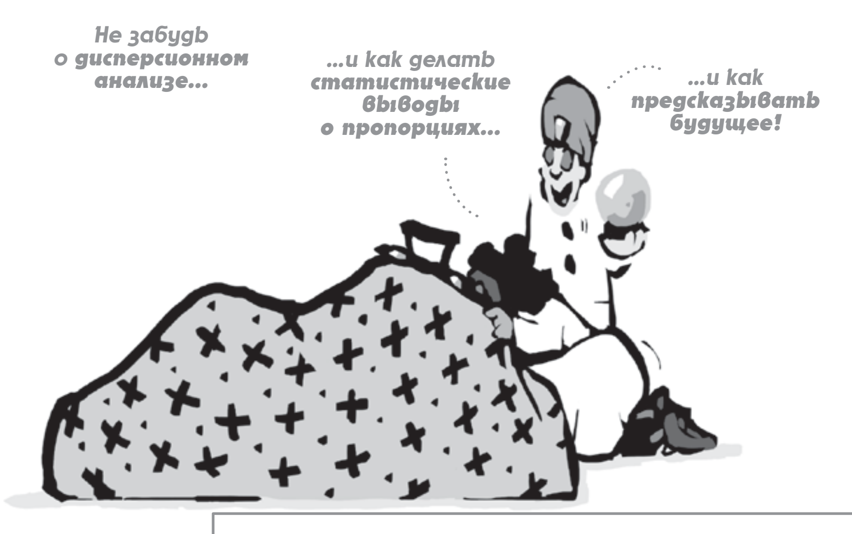




...в нашем распоряжении есть **совершенно другие хитрости.**



Несмотря на тот факт, что **продвинутая статистика** переполнена **разными приемами и хитростями...***



...**основные шаги** формирования статистических выводов **остаются теми же!**

* На стр. **224–225** вы найдете более подробную информацию.

Всегда помните об этом, если захотите углубить свои знания в области статистики.

От нюансов поначалу может голова пойти кругом...



...но, по сути, все проблемы в статистике имеют схожую природу.

И выглядят они вот так:



Разбираться с ними мы будем следующим образом:





Заключение

Мыслить как статистик




На страницах этой книги, так уж вышло, мы рыбачили...
Хотел бы я переловить всех пираний...
...но поймать мы можем только нескольких из них.
...занимались собирательством...


Вот тебе защитный шлем.
Он поможет тебе избежать субъективности.
...и охотились.


Мы никогда не сможем поймать его...
...но мы можем сделать предположение относительно того, сколько их группируется вокруг.
Проделывая все это, мы учились думать как статистики!


Мы не знаем всего...
...но это не значит, что мы не знаем ничего!

В **первой части** мы рассмотрели множества выборочных данных...

Вот засада!

Ну-ка, расскажи нам об **объеме**, **расположении** и **размахе вариаций**.

Все **смещено**!

Вот класс!

...и **исследовали** их.

Осторожнее с **неизвестными переменными**!

Затем, во второй части, мы изучали **статистический вывод** — ...



Мы научились **намечать предполагаемое распределение выборки...**

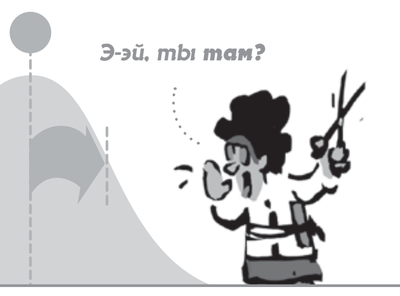
Вот одна банка.

А вот и **план**, как **сто тысяч миллионов банок будут группироваться** в долгосрочной перспективе.

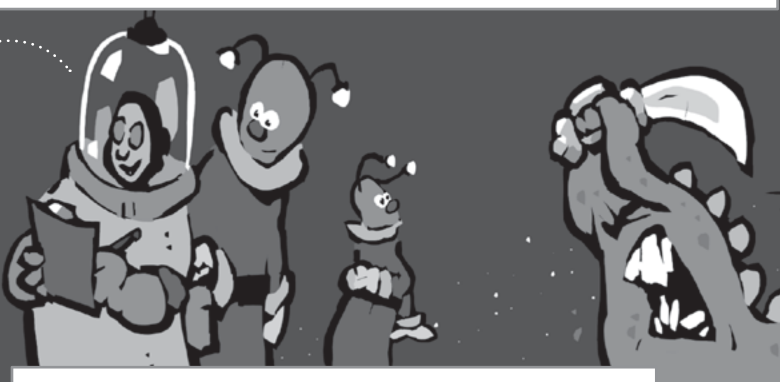
Да это же **центральная предельная теорема**, класс!



...пристально **вглядываться** в него...



...и, **отрезав «лишнее», высчитывать доверительные интервалы...**



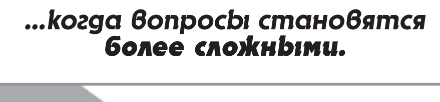
...или **проверять гипотезы.**





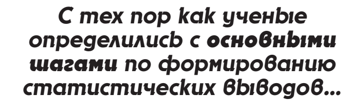
Наконец, мы научились **модифицировать** эти основные шаги...

Бум!



...когда вопросы становятся **более сложными.**

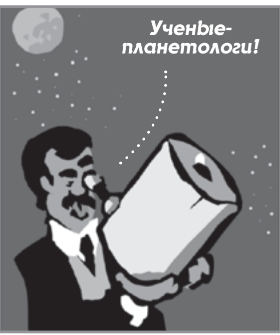






...эти шаги **использовали** самые разные люди...









...при самых **разных обстоятельствах...**





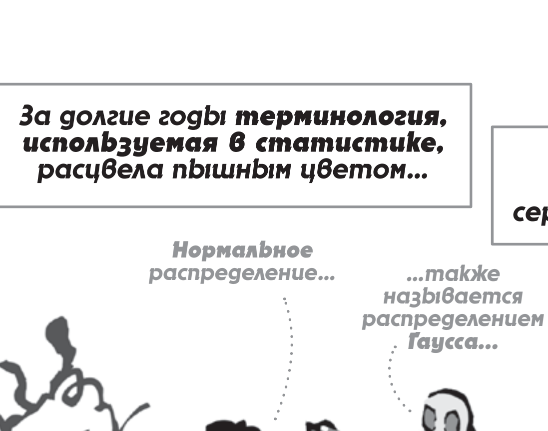




И хотя вышеизложенное и есть суть статистики, обладающей невероятной **силой...**



...она породила немало **загадок.**


За долгие годы терминология, используемая в статистике, расцвела пышным цветом...
...и лексикон статистиков серьезно пополнился...


Нормальное распределение...
...также называется распределением Гаусса...
...а еще z-распределением.
Хм...
...поэтому научиться говорить как статистики...
...может оказаться непростым делом...


Ой, полагаю, эти явления не связаны между собой!
Сколько степеней свободы у остаточной дисперсии?
Чего я ищу: значимости или значения?
А эта ошибка считается стандартной?
...особенно, если вы откроете самый конец книги, где описаны продвинутые методы.


Повторяйте за мной:
P-значение
E-значение
Z- или T-оценка
Критерий хи-квадрат
И это ведь еще не все!
Я уже ни в чем особо не уверен.
А ты уже думал о непараметрическом регрессионном модулировании?

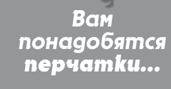






Но если вы хотите научиться **говорить как они...**

...начните обучение...





Приложение

Математическая пещера

Оставь надежду
всяк сюда
входящий...

$$\bar{x} \pm 2\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

...если только
вы не хотите
научиться
говорить
и писать...

...как
настоящий
статистик!

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Когда при изучении выборки мы пишем формулы, то записываем все свои наблюдения таким образом:

$$x_1, x_2, x_3 \ldots x_n$$

где x_1 — это первое наблюдение...

...x_2 — это второе наблюдение...

...а x_n — наше последнее наблюдение в списке, которое заключает в себе n-ное количество наблюдений.

См. стр. 36–37

СЛУЧАЙНАЯ ВЫБОРКА

Формирование выборки произвольным образом — принципиальный вопрос для статистики. Тут важно, чтобы она **не отличалась систематически** от генеральной совокупности, которую представляет.

Выборка, строго говоря, это собрание отдельных наблюдений за конкретными **переменными** (см. ниже). Выборка называется **случайной**, когда состоит из отобранных произвольным образом наблюдений, каждое из которых **независимо** от остальных.

Говоря о случайном отборе данных на страницах этой книги, мы, в частности, имеем в виду **простую случайную выборку (ПСВ)**. Формально **простая случайная выборка (ПСВ)** размера **n** — это собрание n-ного количества наблюдений, полученных таким образом, что **все возможные выборки** n-наблюдений из генеральной совокупности имеют **одинаковую вероятность быть отобранными**.

Иногда срабатывают и другие **неслучайные техники отбора**, например систематический отбор или отбор по стратам (однородным частям в генеральной совокупности — *прим. ред.*), но, какую бы стратегию мы в итоге ни выбрали, мы должны быть уверены, что окончательная выборка будет **представлять всю генеральную совокупность**. Если этого не происходит, наши последующие действия оказываются бессмысленны.

См. стр. 54

ОБЪЕМ ВЫБОРКИ (n)

Это общее количество экспериментов, собранных в одной выборке. В целом, чем больше выборка n, тем больше степень достоверности наших выводов. Но нужно следить за тем, чтобы выборка была сделана **произвольным образом**!

См. стр. 59

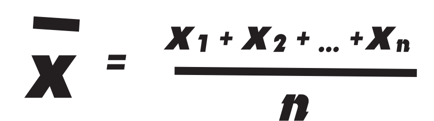
СРЕДНЕЕ ЗНАЧЕНИЕ В ВЫБОРКЕ ($\bar{x}$)

Чтобы высчитать **среднее значение** в выборке, нужно сложить все показатели в ней и разделить их на ее объем. Вот по этой формуле:

Так-с...

...среднее значение в выборке называется «xbar»





$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Среднее значение еще называется **«средним арифметическим»** или просто **«средним»**. На страницах нашей книги мы избегали этого термина, отдавая предпочтение «среднему значению»: нам показалось, что, используя этот более привычный и понятный всем термин, мы облегчим понимание процесса формирования статистических выводов. А еще мы уверены, что большинство наших читателей, услышав слово «среднее арифметическое», все равно тут же подумают о «среднем значении».

Как бы вы его ни называли, среднее значение — самый важный показатель **среднего значения распределения**. Существуют и другие методы, которые могут помочь нам прояснить вопрос формирования определенной совокупности данных, — их выбор всегда зависит от ситуации.

Например, **медиана** — это «центральное значение» в выборке, и в случае смещения оно может быть предпочтительнее. Подобным образом, **усеченное среднее значение** высчитывается путем исключения небольшого процента самых маленьких и самых больших показателей; ему отдают предпочтение, когда в выборке есть экстремальные значения.

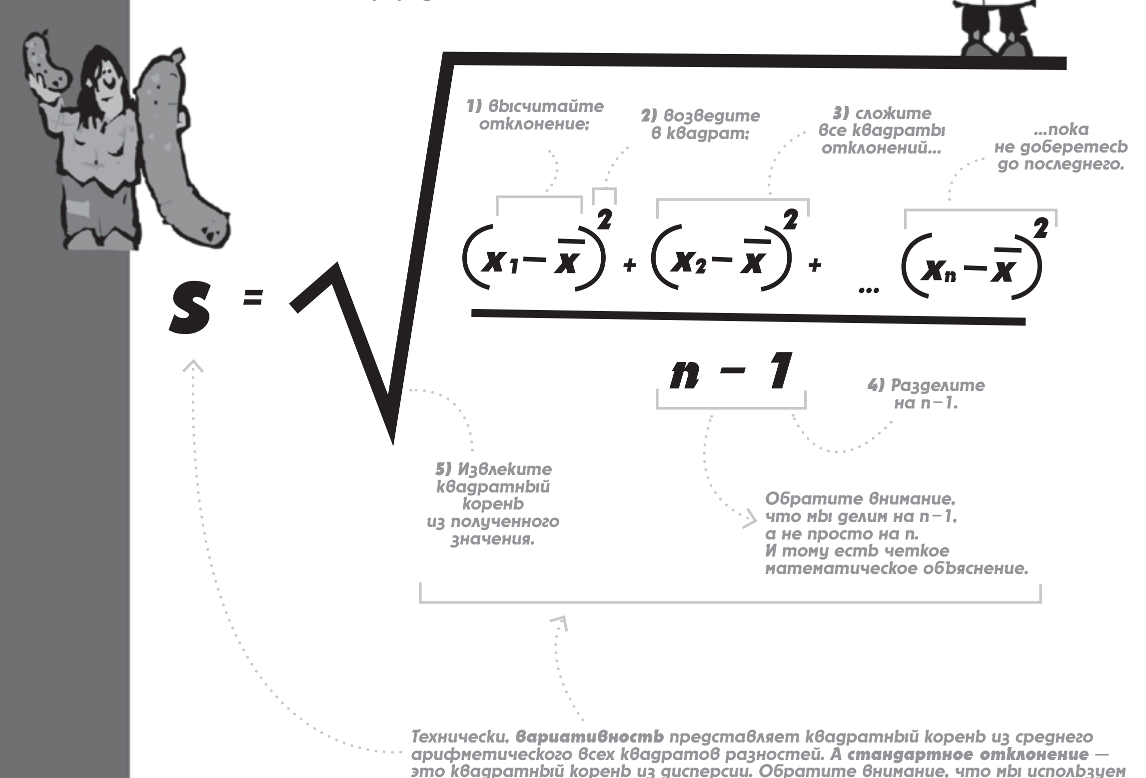
См. стр. 65

СТАНДАРТНОЕ ОТКЛОНЕНИЕ (S)

При подсчете **стандартного отклонения** нам нужно понять, каково среднее расстояние от среднего показателя. Вот как это сделать. Объясняем (буквально) на пальцах:

1) высчитайте расстояние между каждым замером x и средним значением в выборке x. Это и называется отклонение;
2) возведите в квадрат каждое отклонение;
3) сложите все возведенные в квадрат отклонения;
4) разделите сумму на n−1 (остановившись здесь, мы получаем дисперсию);
5) извлеките квадратный корень из полученного значения.

Вот вам формула:



Технически, **вариативность** представляет квадратный корень из среднего арифметического всех квадратов разностей. А **стандартное отклонение** — это квадратный корень из дисперсии. Обратите внимание, что мы используем просто букву s, когда говорим о СО в нашей выборке.

См. стр. 70

ПЕРЕМЕННАЯ (X)

Переменная — то, что нас особенно интересует. Но так как в статистике мы всегда собираем данные произвольным образом, то называем те переменные, которые ищем, **случайными**. По определению, **случайная переменная** — это переменная, значение которой совершенно случайно.

В краткосрочном периоде мы не имеем никакой возможности предсказать значение случайной переменной, пока не получим ее саму (как, например, в случае подбрасывания монеты). В долгосрочном периоде мы предсказываем значение случайной переменной, используя **вероятность** (см. ниже).



См. стр. 84

РАСПРЕДЕЛЕНИЕ

Говоря математическим языком, **распределение** описывает расположение всех возможных показателей случайной переменной. Если, например, вы создали гистограмму со всеми показателями переменной генеральной совокупности, у вас получится **распределение генеральной совокупности** для этой переменной.

Если говорить более общо, распределения позволяют нам высчитать **вероятности** случайных показателей из конкретного интервала. Делая статистические выводы, мы высчитываем вероятности, используя **распределение выборки** (см. ниже). Но если бы мы имели на руках распределение генеральной совокупности, мы могли бы использовать и его для подсчета вероятности. Вот как это делается.

Если бы мы знали, каково распределение всей генеральной совокупности рыбы в озере, отсортированной по длине...

...мы могли бы подсчитать процентное соотношение рыбы в любой области этого распределения...

...например, **эта область** охватывает интервал длиной 8–12 см.

Отсюда следует, что, если бы мы нырнули в озеро и **случайным образом выбрали одну рыбешку...**

...**вероятность** того, что ее длина колебалась бы в пределах между 8 и 12 см, была бы такой же (в процентном соотношении), как и область распределения внутри затемненной части.

Если половина всех рыб имеет длину, колеблющуюся в пределах между **8** и **12** см...

...то вероятность того, что я случайно поймаю рыбку такой длины, равняется **50%**, или **0,5**.

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Конечно, на деле мы никогда не можем посмотреть на распределение всей генеральной совокупности. Если бы мы могли, нам бы не была нужна статистика.

См. стр. 85

СТАТИСТИЧЕСКИЕ ДАННЫЕ ВЫБОРКИ ПРОТИВ ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Раз уж в статистике для суждения о **генеральной совокупности** используются выборки, для каждой из них есть свои термины и технические примечания.

Характеристики выборки называются **статистическими данными.**

Когда мы выводим формулы, **xbar** подразумевает только **среднее значение в выборке.**

$\bar{x}$

S относится исключительно к стандартному отклонению в выборке.

s

Статистические данные — это то, что мы, собственно, изучаем, и потому судить о них можем со всей точностью.

Характеристики генеральной совокупности называются **параметрами.**

Греческая буква нижнего регистра **мю** относится исключительно к **среднему значению в генеральной совокупности.**



Греческая буква нижнего регистра **сигма** относится исключительно к **стандартному отклонению в генеральной совокупности.**



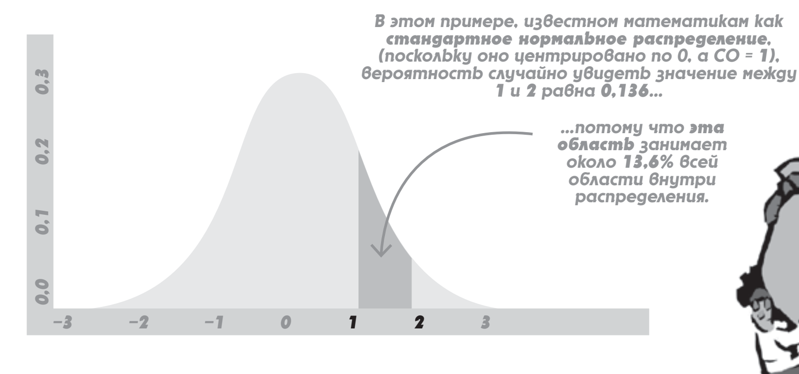
Параметры — это то, что нас интересует в итоге, но о чем мы можем только строить предположения.

См. стр. 94

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

В математике и в теории вероятностей существует огромное количество распределений, которые бывают самых разных форм. Наиболее известно **нормальное**. В статистике оно особенно важно, потому что показывает, как группируются средние значения (см. «**ЦПТ**» ниже)

Как и в случае с любым другим распределением, мы всегда можем поделить нормальное распределение на области, которые будут отображать вероятности внутри него. Мы говорили о том, как это сделать, на стр. **115**. Но вот еще пример.



См. стр. 95

РАСПРЕДЕЛЕНИЕ ВЫБОРКИ

Говоря технически, это распределение статистической выборки. И хотя мы можем наметить распределение выборки для любого вида статистики (СО, медианы и т. д.), мы здесь фокусируемся на распределении выборки, сделанном на базе средних значений. Так, например, если бы мы отобрали из генеральной совокупности много-много выборок размера n, высчитали бы $\bar{x}$ для каждой из них, затем нарисовали гистограмму всех $\bar{x}$ значений, то получили бы распределение выборки $\bar{x}$. Примером может послужить сарайчик Безумного Билли, где он хранит снасти (см. стр. 107). Распределение выборки послужит ключом к статистическим выводам.

См. стр. 101–102

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА (ЦПТ)

Многие статистические выводы зависят от **ЦПТ**, которая утверждает, что распределение выборки $\bar{x}$ становится приближенно нормальным, по мере того как увеличивается объем выборки **n**.

Если говорить более детально, то для **случайных выборок** большого объема **n**, отобранных из одной генеральной совокупности со средним значением **μ** и **СО σ**, распределение $\bar{x}$ будет приближенно нормальным со средним значением **μ** и СО, равным **σ, деленной на корень квадратный из n**.



ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА (ПРОДОЛЖЕНИЕ)

ЦПТ — это очень общий результат, который почти всегда будет применяться так, как описано в этой книге. Однако есть несколько **важных условий**, лежащих в основе ЦПТ.

Во-первых, ЦПТ работает только в том случае, если все показатели x, x_1, x_2, x_3 ... x_n в нашей выборке взяты из **того же самого распределения генеральной совокупности**. Это обычно касается выборок, полученных на практике, но может быть важно, если мы разбираемся и с более сложными вопросами.

Во-вторых, каждый замер x_i должен быть **случайным**. Строго говоря, это означает, что все показатели x_i должны быть **независимы** друг от друга. Например, замеры температуры, сделанные в каком-то географическом регионе, не будут независимы, поскольку результаты измерений из одного конкретного места будут очень похожи на замеры температуры в другом месте поблизости. Статистики скажут, что они **«коррелируют друг с другом»**, потому что существует некий общий принцип, оказывающий влияние на значение каждого из x_i (см. **«Корреляция»** ниже).

Наконец, ЦПТ применима, когда **n** стремится к бесконечности, однако на практике мы используем приближенную версию ЦПТ, которая работает, когда $n \geq 30$. Мы называем такие значения n «большими». Это звучит довольно условно, но, если вдаваться в подробности, мы погрязнем в математике.

См. стр. 112

ВЕРОЯТНОСТИ

В книге мы обозначаем вероятности процентами (например, 95%), но в математике для тех же целей используют числа от **0** до **1** (например, **95% = 0,95**). Так, формально **вероятность** — это число между **0** и **1**, которое отображает степень возможности наступления какого-то события. Однако вся сложность заключается в том, что ее высчитывают **только в долгосрочной перспективе**.

Если, например, среди избирателей одинаковое количество мужчин и женщин, вероятность того, что случайно отобранный избиратель окажется женщиной, будет равна **0,5**. Но первые несколько избирателей, отобранных произвольным образом, вполне могут оказаться мужчинами чисто случайно. То есть показатель **0,5** относится к **долгосрочной перспективе:** если мы произвольным образом отберем достаточно избирателей, **в конечном счете** получим равное количество мужчин и женщин.

В другом примере, когда мы подбрасываем монетку, с вероятностью 0,5 выпадет либо орел, либо решка. Но если мы подбросили монетку только один раз и выпал орел, вероятность того, что в следующий раз будет то же, равна **по-прежнему 0,5**. В этом случае каждый бросок монетки **независим** от другого.

Подводя итог, скажем, что всякий раз, когда мы высчитываем вероятность, она может быть выражена числом между 0 и **1** (**или 0 и 100%**), и это число всегда соотносится с областью внутри вероятностного распределения. По определению, вся область внутри него равна **1**.



См. стр. 114

ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТИ

Мы можем высчитать области внутри любого распределения (например, нормального, как мы видели на картинке на стр. **114**), используя интегрирование, которое представляет собой **вычислительный метод**. На практике все необходимые вычисления делаются на компьютере.

В примере с сарайчиком Билли распределение выборки имеет нормальную форму из-за ЦПТ. Однако в огромном количестве других сфер применения статистики какое-то конкретное распределение выборки может и не иметь нормальной формы, но мы по-прежнему сможем сделать все необходимые вычисления. Если хотите подробностей, вернитесь к главе 14.

Все распределения можно изобразить в качестве кривых, но также они могут быть представлены **функциями**. Их принцип похож на принцип работы компьютера, где есть **исходные данные** (в нашем случае случайные переменные) и **результат** (в нашем случае **вероятность**).

Среди условных знаков есть обозначение, принятое для функции вероятности f со средним значением μ и СО σ:

> Если X представляет собой дискретную случайную переменную с распределением f, **зависящим от** μ и σ...
>
> ...то f, **зависящее от** μ и σ **(x)**, равно вероятности того, что X примет значение x.

К сожалению, дальше все только еще больше усложняется и запутывается. Вот, например, функция вероятности для нормального распределения:

$$h_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



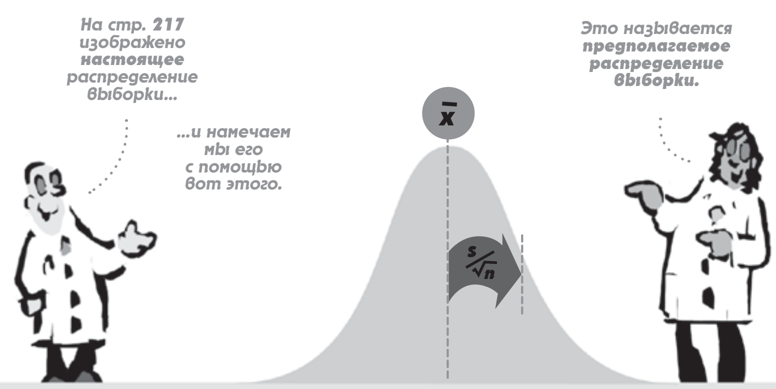
Хотя на первый взгляд она выглядит **устрашающе**, в целом функции вероятности, подобные этой, оказываются необычайно полезными, потому что они связывают **определенные виды случайных событий** (например, попытки поймать рыбу определенного размера) с **предсказуемыми долгосрочными выводами** (как часто вы сможете рассчитывать на это в долгосрочной перспективе).

См. стр. 129

ОЦЕНКА РАСПРЕДЕЛЕНИЯ ВЫБОРКИ

На практике, когда мы используем ЦПТ, мы никак не можем узнать настоящие значения параметров μ и σ, поэтому используем статистику $\bar{x}$ и **S**, чтобы **приблизить** их **друг к другу**. **Аппроксимация** работает потому, что мы собираем данные произвольным образом. В результате мы ожидаем, что $\bar{x}$ будет отличаться от μ, а **S** — от σ, но **только из-за случайной вариации**.

После того как мы заменили приближенные значения, мы можем называть получившийся результат **предполагаемым распределением выборки**.



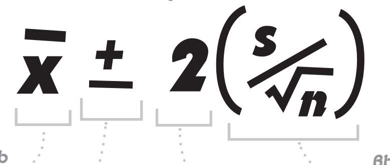
Обратите внимание, что мы можем построить **предполагаемое распределение выборки** и для других статистических данных, например **S** (см. стр. **201**, история про петарды). Однако мы можем ждать, что распределение выборки будет нормальной формы, только если применим ЦПТ или похожие результаты.

См. стр. 135

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

*Строго говоря, это такой тип интервальной оценки, который отражает определенную **степень достоверности**. Доверительный интервал можно высчитать для любого параметра, хотя какие-то специфические технические детали могут меняться. Вот формула, по которой можно высчитать девяностопятипроцентный **доверительный интервал** для среднего значения в генеральной совокупности **μ**:*

$$\bar{x} \pm 2\left(\frac{s}{\sqrt{n}}\right)$$



Вот какое заключение мы можем сделать из этой формулы.

Мы на 95% уверены...

*...что **μ** находится где-то в этом пределе.*

Предполагаемое распределение выборки для $\bar{x}$

*Мы можем **изменить степень достоверности, передвинув отсечку**. Например, если бы нам был нужен доверительный интервал в **80%** для среднего значения в генеральной совокупности, наша отсечка была бы на отметке **1,3**, потому что приблизительно 80% нормального распределения имеет **1,3** СО относительно центра (см. пример на стр. **157**).*

*В идеале нам нужен **максимально узкий интервал** для любой степени достоверности, поскольку он точнее. Одним из безошибочных способов сузить интервал будет увеличение **n** (для этого нужно собрать больше материала для наблюдений). Вот почему чем больше выборка, тем лучше! (см. пример на стр. **159**).*

*Помните, что **степень достоверности** зависит от значения вероятности. Поэтому все это имеет смысл, только когда мы думаем о долгосрочной перспективе. В результате, высчитывая интервал с помощью формулы, приведенной выше, мы даже не знаем, **учитывает ли она μ или нет**! Все, что мы можем сказать, это то, что интервалы, намеченные таким образом, будут точными в долгосрочной перспективе. В случае с **95%-ным** доверительным интервалом мы понимаем, что наши заблуждения составят **5%** случаев... в долгосрочной перспективе.*

См. стр. 163

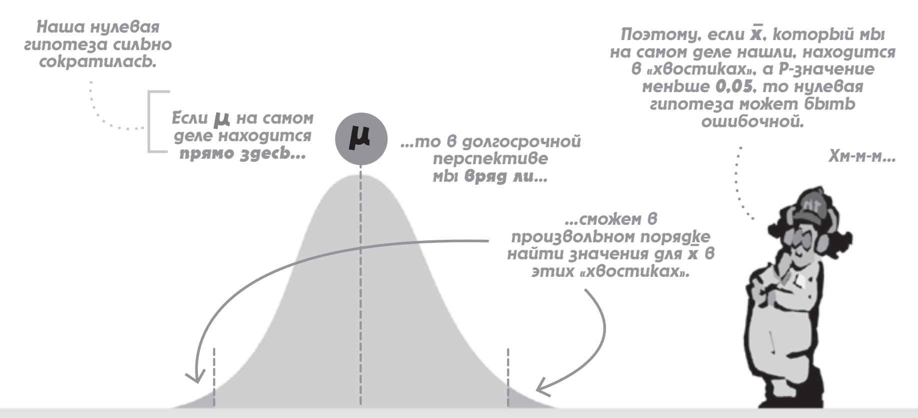
ПРОВЕРКА ГИПОТЕЗЫ

*При проверке гипотез используются те же самые **статистические методы**, что и при вычислении доверительных интервалов. Начинаем мы с построения предполагаемого распределения выборки. Но на этот раз оно нам нужно, чтобы ответить на вопрос: истинно или нет какое-то **конкретное значение** для параметра генеральной совокупности. И делаем мы это, проверяя, **насколько совместимы** исследуемые нами данные с тем конкретным значением.*

*Формально проверка начинается с рассмотрения двух гипотез: нашей исследуемой (иногда ее называют **альтернативной**) и **нулевой** (в книге мы использовали слова «скучная», «неоригинальная», «обыкновенная»).*

*Проверка гипотезы заканчивается тем, что мы высчитываем **Р-значение** и используем его, чтобы вынести **формальное решение**, находятся ли наши данные достаточно далеко от параметра, предсказанного нулевой гипотезой, чтобы оправдать выбор, сделанный в пользу иного объяснения.*

Вот краткий комментарий к описанному выше.



Предполагаемое распределение выборки для $\bar{x}$, при условии, что нулевая гипотеза верна

В этой книге мы проверяем гипотезы, касающиеся средних значений. На практике эти же общие принципы применимы к любому параметру и соответствующей статистике — отличаться будут только математические данные.

См. стр. 169

Р-ЗНАЧЕНИЕ

*Формально **Р-значение** — это вероятность ошибки при отклонении нулевой гипотезы. В книге мы обозначали Р-значение процентами, хотя часто бывает и так, что его обозначают числами **0–1**. Р-значение, равное **5%**, часто выражается числом **0,5**.*

*Иногда мы вычисляем Р-значение для обоих «хвостиков» нашего предполагаемого распределения выборки (см. стр. **157**, это называется **«двусторонний тест»**), **а иногда мы вычисляем Р-значение только для одного из них (см. стр. 159, это называется «односторонний тест»)**. Какой из них выбрать, зависит от того, какого рода исследуемая гипотеза нас интересует.*



*На практике для **вычисления Р-значения** пользуются компьютерами. Важно отметить, что (когда мы проводим двусторонний тест) вероятность **0,5** — это именно та область, которая не помещается внутри **95%-ного** доверительного интервала. Можно прибегнуть к сравнительно простому способу проверки гипотезы: наметить **95%-ный** доверительный интервал для **μ**, как описано выше. Если значение **μ**, предсказанное нулевой гипотезой, не находится в этом интервале, значит, Р-значение меньше 0,05.*

*В подобных случаях увеличение n приводит к уменьшению Р-значения. Поэтому, по мнению статистиков, сбор большего количества материала для исследования — это безошибочный способ с полным на то **правом** отклонить нулевую гипотезу. Это еще раз доказывает, что чем больше объем выборки, тем лучше!*

P-ЗНАЧЕНИЯ (ПРОДОЛЖЕНИЕ)

Помните, что P-значение измеряет вероятность, поэтому оно актуально, только когда мы думаем о долгосрочной перспективе.

На практике мы отклоняем нулевую гипотезу, если наше P-значение «достаточно мало», что (согласно общепризнанному правилу) означает ***меньше 0,05****. Но в этом числе нет ничего магического. Вероятность «меньше чем 0,05» означает то же, что и «меньше 1 из каждого 20 случая в долгосрочной перспективе».*

Так, например, если мы проверяем гипотезу и у нас получается P-значение, равное 0,049, это означает, что «если нулевая гипотеза была бы истинной, мы бы чисто случайно видели данные, подобные нашим, примерно 49 раз из каждой ***1000*** *в долгосрочной перспективе». Так как 49/1000 меньше, чем 1/20, мы бы решили, что наши данные не очень хорошо соответствуют нулевой гипотезе.*



См. стр. 172

МЫ МОЖЕМ ОШИБАТЬСЯ



В основе ***всех*** *статистических исследований лежит случайная выборка, а в основе всех статистических выводов лежит вычисление вероятности. В результате всякий раз, когда мы используем статистические данные малых выборок, чтобы сделать предположение о каком-нибудь параметре в генеральной совокупности, мы* ***можем ошибаться!***

Из-за этого нам нужно ***очень внимательно*** *следить за языком, ведь нас так и подмывает порой сделать* ***громкое заявление****, опираясь на статистические выводы. Нам нужно быть особенно осторожными, когда мы делаем формальные выводы, основываясь на P-значениях. Потому что мы используем P-значения только тогда, когда исследуем теории, которые нам особенно по душе.*

Если мы проверяем теорию и используем маленькое P-значение, чтобы подтвердить ее, есть вероятность, что ***мы заблуждаемся****. Наша теория может оказаться ошибочной, и тогда наши результаты лучше объяснить случайной вариацией. В статистике это называется ошибкой* ***первого рода****.*

Если же мы проверяем теорию с использованием больших P-значений с целью отклонить ее, ***мы можем ошибаться****, а наша теория — быть верной, и мы получим результаты, близкие по значению к нашей нулевой гипотезе. В статистике это называется ошибкой* ***второго рода****.*

Резюмируя, скажем, что проверка гипотезы сводится к поиску ответа на вопрос «Какова вероятность того, что мы получили результаты совершенно случайно?» Их ведь нельзя использовать ни для убедительного опровержения, ни для убедительного доказательства теории. Они нужны ***только*** *для того, чтобы помочь нам* ***дискредитировать нулевую гипотезу****.*

В статистике, так уже повелось, мы в любую секунду ***можем оказаться неправы****. И так оно всегда и бывает. А все потому, что мы привязаны к долгосрочности, когда оцениваем краткосрочные наблюдения.*



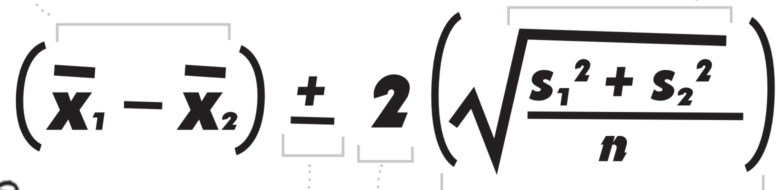
См. стр. 197

СТАТИСТИЧЕСКИЕ ВЫВОДЫ О РАЗНИЦЕ

Чтобы высчитать доверительный интервал, касающийся разницы между двумя средними значениями в генеральной совокупности, мы можем использовать формулу, которая совсем немного отличается от той, что мы описали выше.

В этом случае нас интересует разница между двумя средними значениями генеральной совокупности, и мы высчитываем ее с помощью двух средних значений в выборке.

Вот так мы объединяем вариации двух генеральных совокупностей.



$$\left(\bar{x}_1 - \bar{x}_2\right) \pm 2\left(\sqrt{\frac{S_1^2 + S_2^2}{n}}\right)$$



Плюс и минус означают, что мы сдвинулись в сторону от середины в обоих направлениях.

Мы используем **отсечку**, представленную цифрой 2, если хотим 95%-ный доверительный интервал. Но, конечно, допустимы любые изменения.

Все это — **СО нашего предполагаемого распределения выборки**. В этом случае оно равно 1.

В нашем случае $\bar{x}_1$ = 59,7, S_1 = 4,6, $\bar{x}_2$ = 44,2, а S_2 = 4,7.

Необходимо отметить, что эта формула может иметь другой вид. Например, она изменится, если мы используем **разные** объемы выборки или если они **слишком маленькие, чтобы получилось нормальное распределение.**

См. стр. 199

СТАТИСТИЧЕСКИЙ ВЫВОД О ВЫБОРКЕ НЕБОЛЬШОГО ОБЪЕМА

Когда бы мы ни формировали статистические выводы о среднем значении генеральной совокупности при наличии **выборки маленького объема (например, когда n меньше 30)**, мы не можем положиться на ЦПТ. В таком случае нам понадобится то, что известно как ***t*-распределение**. Оно работает, только когда сама генеральная совокупность нормальная.

По каким-то не вполне понятным историческим причинам, связанным с тем фактом, что этот вид распределения был открыт и впервые описан парнем, работавшим на пивоварне Guinness, ***t*-распределение** еще называют **распределением Стьюдента**.

Оно шире нормального распределения и на практике видоизменяется в зависимости от **n** (чем меньше **n**, тем шире **t**). В результате, когда мы используем *t*-распределение вместо нормального, наши доверительные пределы становятся шире, а наше Р-значение — больше. В обоих случаях нам требуются дополнительные статистические данные, чтобы добиться той же степени достоверности. Как будто нас наказывают за небольшой объем выборки.

В нашем случае специфическое распределение, которое мы использовали, и представляет собой пример t_9, которое называют **«t с 9 [как в случае с n − 1] степенями свободы»**.

А вот и замеры уровня РН, которые мы делали у 10 произвольным образом отобранных пришельцев: **2,09; 2,39; 1,32; 2,99; 2,62; 2,60; 2,45; 2,13; 2,27** и **2,95**. Имея эти данные, мы можем высчитать, что **$\bar{x}$ = 2,38, S = 0,48**, а **n = 10**. Затем мы вставляем все эти значения в следующую формулу и получаем **95%-ный** доверительный интервал.

***t*-распределения** выглядят почти нормально. Единственное, они немного шире вот в этих частях:

$$\bar{x} \pm 2{,}26\left(\frac{s}{\sqrt{n}}\right)$$

Мы используем другую **отсечку** в зависимости от того, какого рода у нас *t*-распределение, и от того, насколько хотим быть уверены.





См. стр. 201

СТАТИСТИЧЕСКИЕ ВЫВОДЫ О СТАНДАРТНОМ ОТКЛОНЕНИИ



*В истории с маленькой хулиганкой Сьюзи мы не можем рассчитывать на ЦПТ. Распределение выборки со стандартным отклонением **не обязательно** будет нормальным, и при вычислении доверительного интервала мы **не можем** пользоваться ничем, похожим на формулу, приведенную на стр. **220**. Но **основные шаги по-прежнему те же:** мы создаем определенный вид распределения выборки (при помощи каких-нибудь эффектных математических вычислений), намечаем **95%-ный** доверительный интервал и высчитываем все вероятности в этом пределе (опять же, пустившись в сложные подсчеты).*

*Вся эта эффектная математика слишком запутанна, чтобы объяснять ее здесь, но в основе всех вычислений все равно лежат показатели задержки срабатывания 15 случайным образом отобранных петард Dingalings: **2,05; 2,25; 2,33; 2,40; 1,66; 2,39; 1,89; 2,18; 2,06; 1,89; 2,14; 2,38; 2,07**. Проанализировав эти данные, получаем $\bar{x}$ **= 2,14**, **S = 0,21** и **n = 15**.*

*Обратите внимание: чтобы сформулировать статистические выводы, нам **нужно предположить, что генеральная совокупность нормальна**. В этом случае предположение может быть справедливым, но если мы имеем дело с каким-то дефектом производства, который смещает всю генеральную совокупность (например, неожиданно большая задержка срабатывания у петард с одной фабрики может быть сбоем, а при анализе этого не учтут), тогда наши выводы **могут оказаться ошибочными**.*

*Так же, как и t-распределение, распределение выборки со стандартным отклонением меняется в зависимости от объема выборки. И, как и во всех случаях, чем больше **n**, **тем более мы уверены в наших выводах**.*

См. стр. 202

КОРРЕЛЯЦИЯ

***Корреляция** искажает огромное количество статистических анализов. Если замеры нашей выборки $x_1, x_2, x_3 \ldots x_n$ «коррелируют» друг с другом, они не будут независимыми и мы не можем применить к ним ни одну из наших методик. Поэтому в случае, когда данные коррелируют друг с другом, при формировании статистических выводов нам необходимо принимать этот факт во внимание.*

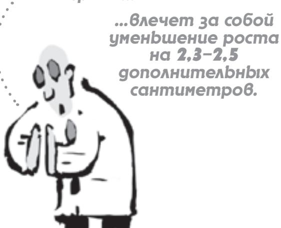
*Например, при исследовании состояния здоровья двух **биологических близнецов** следует учитывать тот факт, что все данные об одном из них непременно коррелируют с данными брата или сестры, но больше ни с чьими. Эту корреляцию можно уменьшить, применив **«парный тест»**. Другие исследования предполагают замеры, коррелирующие друг с другом по **географическому принципу** (как было в примере с гекконами на стр. **202**) или **во времени** (представьте себе случай, когда PH-уровень слюны одного пришельца варьируется в зависимости от времени суток и за отчетный период мы получаем огромное количество замеров у одного и того же инопланетянина). Иногда мы можем внедрить этот тип корреляции, предвосхищая его в более крупных математических моделях. Хотя, строго говоря, каждый вид корреляции предполагает применение своих трюков.*

См. стр. 203

РЕГРЕССИОННЫЙ АНАЛИЗ

*Он необходим, когда нужно исследовать природу взаимоотношений зависимой **переменной** и одной или нескольких **независимых переменных**. В примере на стр. **203** лекарство, заставляющее выпившего уменьшиться в размерах, будет независимой переменной, а то, насколько он уменьшается, — переменной зависимой.*





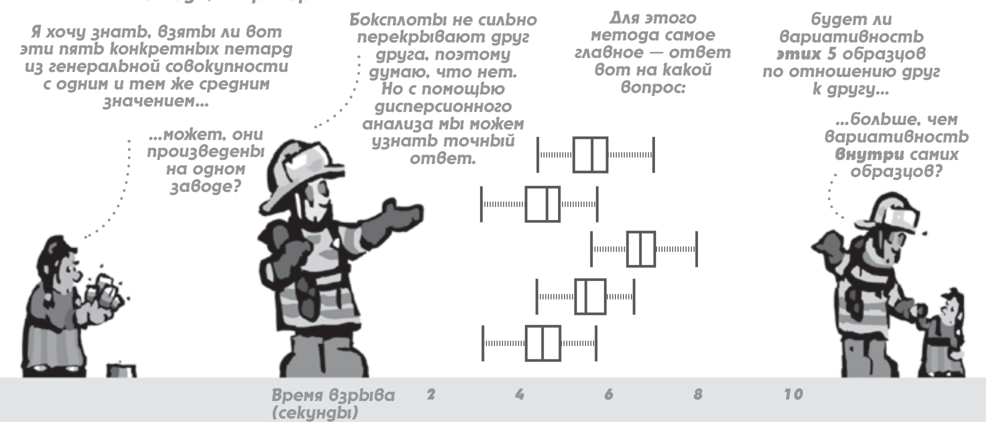
*На практике нам нужно быть предельно внимательными, когда мы проявляем любопытство относительно того, могут ли независимые переменные на самом деле **привести** к изменениям в переменных зависимых. Причинно-следственная связь — вещь очень тонкая.*

*Мы также можем использовать этот тип анализа, чтобы сравнивать характеристики, например **вес** или **рост**, в генеральной совокупности. В таком случае наклон прямой скажет нам об изменениях в весе, происходящих по мере увеличения роста на каждый сантиметр.*

См. стр. 203

ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ представляет собой технику проверки гипотез, но она сильно отличается от того способа, который мы рассматривали на страницах этой книги. Эта техника построена на сравнении **вариантов между группами**, а также **внутри них**. Существует множество способов применения этого метода, например:



См. стр. 203

СТАТИСТИЧЕСКИЕ ВЫВОДЫ О ПРОПОРЦИЯХ



Если нам интересно, **какова пропорция** футбольных фанатов, предпочитающих сырные шарики свиным ребрышкам, и **каков процент** избирателей, которые с большой долей вероятности отдадут голоса за переизбрание сенатора Сэма Ворма на предстоящих выборах, мы можем использовать методы формирования статистических выводов, о которых говорили в этой книге, но нам нужно уточнить детали.

Например, таким образом мы можем высчитать 95%-ный доверительный интервал для пропорции совокупности **p**.

Мы берем нашу пропорцию выборки $\hat{p}$, чтобы оценить пропорцию генеральной совокупности.

Это стандартное отклонение пропорции выборки.

$$\hat{p} \pm 2\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$



Имея выборку большого объема, мы можем использовать стандартное нормальное распределение. Итак, чтобы получить вероятность, равную 95%, мы используем 2 отсечки.

Еще раз, нам нужна выборка большого объема, чтобы все получилось, и чем больше будет объем, тем уже будет наш интервал!

В опросах это называется **предел погрешности.**

См. стр. 203

ПРЕДСКАЗАНИЕ БУДУЩЕГО



В нашей книге мы фокусировались на использовании данных выборки для того, чтобы проанализировать генеральную совокупность, ее среднее значение или стандартное отклонение. Но мы также можем использовать статистические выводы, чтобы сделать предсказания о каких-то **отдельных наблюдениях**. Например, мы можем задавать себе вопросы типа этого: «Основываясь на имеющихся у меня замерах, можно предположить, каким, вероятнее всего, будет **следующий замер** ($x_n + 1$)?»

Опять-таки, мы можем продвинуться немного вперед, предприняв основные шаги, необходимые при формировании статистических выводов. Например, если предположить, что генеральная совокупность нормальная (всегда опасное предположение), мы можем создать **«интервал прогноза»**, который будет похож на стандартный доверительный интервал, но окажется немного шире.

На практике это используется, чтобы предсказать погоду или цены на финансовых рынках, хотя там применяют более сложные техники.



Об авторах

Грейди Клейн — художник-мультипликатор, карикатурист и иллюстратор, соавтор книг «Микроэкономика. Краткий курс в комиксах» и «Макроэкономика. Краткий курс в комиксах», а также создатель серии графических новелл «Затерянная колония». Живет в Принстоне (США) с женой и двумя детьми.

Алан Дебни — доктор философии, отмеченный многочисленными наградами адъюнкт-профессор статистики в Техасском Университете A&M. Живет в Колледж-Стейшен (США) с женой и тремя детьми.

Максимально полезные книги от издательства «Манн, Иванов и Фербер»

Научно-популярное издание

The Cartoon Introduction to Statistics
By Grady Klein and Alan Dabney, Ph. D.

Грейди **Клейн**
Алан **Дебни**

Статистика

Базовый курс в комиксах

Главный редактор *Артем Степанов*
Ответственный редактор *Ольга Киселева*
Научный редактор *Ирина Николаева*
Литературный редактор *Анна Санникова*
Арт-директор *Алексей Богомолов*
Дизайн обложки *Сергей Хозин*
Верстка *Людмила Гроздова*
Корректоры *Мария Кантурова, Надежда Болотина*

НАУКА В СЛОВАХ И КАРТИНКАХ

Грейди Клейн — графический дизайнер, аниматор и художник комиксов из Принстона. Автор графических романов и соавтор нескольких научно-популярных комиксов.

Алан Дебни — профессор статистики Техасского университета A&M, исследователь больших данных, лауреат нескольких профессиональных премий за креативные решения в преподавании.

Комикс с драконами, великанами и инопланетянами в главных ролях — на самом деле неожиданно захватывающее и доступное пособие по статистике. Его создатели — профессор статистики Алан Дебни и художник Грейди Клейн — нескучно объясняют, как собирать надежные данные, делать правильные выводы, владея ограниченной информацией, оценивать результаты опросов и обращаться с множеством цифр, окружающих нас. Вы получите базовые знания о сложной науке самым увлекательным способом. А если, пройдя этот квест, захотите углубиться в изучение статистики, в конце книги вас ждет «Математическая пещера», полная формул и пояснений. Эта книга — пожалуй, лучший путеводитель по миру, которым правят данные.



Эта книга — лекарство для всех, кто страдает фобией статистики.

Чарльз Уилан, автор книги «Голая статистика. Самая интересная книга о самой скучной науке»

Слава богу, кто-то умудрился написать книгу о статистике, которую весело читать. И будьте осторожны: возможно, вы не захотите с ней расстаться, пока не дочитаете до конца.

Себастьян Трун, крутой парень из Google и CEO образовательной ассоциации Udacity

Это большой-большой секрет, что статистика — штука веселая, имеет отношение к каждому из нас и способствует развитию интеллекта. Грейди Клейн и Алан Дебни разболтали его, отправив нас в путешествие по фундаментальным идеям, делающим статистику важной частью нашей жизни, в которой полным-полно самых разных фактов.

Джон Стори, профессор геномики и статистики в Принстонском университете

ISBN 978-5-00100-260-4



9 785001 002604 >

Максимально полезные книги на сайте **mann-ivanov-ferber.ru**

издательство
МАНН, ИВАНОВ И ФЕРБЕР

Like facebook.com/**mifbooks**

vk.com/**mifbooks**

instagram.com/**mifbooks**